

NONPARAMETRIC LINEAR REGRESSION FOR SPATIAL DATA ON GRAPHS WITH WAVELETS

JOHANNES THEODOR NIKOLAUS KREBS

ABSTRACT. Nonparametric regression estimates for d -dimensional data are studied. The data is defined on a not necessarily regular N -dimensional lattice structure and is strong mixing. We show the consistency and get rates of convergence for nonparametric regression estimators which are derived from finite dimensional linear function spaces. As an application, we choose linear spaces which are spanned by d -dimensional wavelets. Furthermore, we give numerical applications of the developed theory.

INTRODUCTION

Nonparametric regression is a well established technique in statistics. Classical textbooks investigate the properties of nonparametric regression estimates under the assumption that the sample data is independent and identically distributed. In many statistical and real world situations however, sample data is not independently distributed. A typical example is data which is observed on a spatial structure such as a regular N -dimensional lattice or more generally a graph $G = (V, E)$. Especially in the area of Markov random fields it is convenient to assume the dependency structure of the data to be determined by the adjacency matrix of the graph and to vanish with increasing graph distance. A particular application which we have in mind are data like traffic intensity or road roughness indices on road networks which may be represented as graphs.

We consider for a given strong mixing random field $\{(X(s), Y(s)) : s \in V\}$ with equal marginal distributions the regression model $Y(s) = m(X(s)) + \varsigma(X(s))\varepsilon(s)$ where m and ς are two elements from the function space $L^2(\mu_X)$ with the notation that μ_X is the distribution of $X(s)$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The error terms $\varepsilon(s)$ are $(0, 1)$ distributed and independent of X . We aim at estimating m with the truncated least-squares estimator as defined in Györfi et al. [2002]. We give sufficient conditions both on the data and on the estimator such that the nonparametric truncated least-squares routine is consistent. Furthermore, we state results on the rate of convergence of our procedure. As examples of application for estimating the conditional mean function m , we use a d -dimensional wavelet approach to construct a dense subspace of $L^2(\mu_X)$ in which we define the empirical estimator \hat{m}_n of the function m .

For the numerical part, we consider random fields which are simulated on graphs with an ansatz based on the concept of cliques which is due to Kaiser et al. [2012]. This approach puts us in position to consider our simulation as iterations of an ergodic Markov chain. We visualize this theory in two simulation examples where we consider one bivariate and one univariate nonparametric linear regression problem.

This paper is organized as follows: we give in Section 1 the basic notions which we use throughout the paper. Furthermore, we state two general theorems on the consistency and the rate of convergence of the nonparametric truncated linear least-squares estimator. In Section 2 we show how general d -dimensional wavelets fulfill the requirements for the consistency and rate of convergence statements of the previous section. The last Section 3 is devoted to numerical applications: we give simulation concepts for random fields that are defined on an arbitrary graphical structure and discuss the developed theory in two examples. Section 4 contains the proofs of the presented theorems. Appendix A contains useful exponential inequalities for dependent sums. Appendix B, contains a piece of ergodic theory for spatial processes.

1. LINEAR REGRESSION ON STRONG SPATIAL MIXING DATA

In this section, we present two main results of this article. We focus on random variables that are defined on a spatial structure, in particular an N -dimensional lattice. We make a some definitions

Date: September 25, 2016.

2010 Mathematics Subject Classification. Primary 62G08, 62H11, 65T60; secondary: 65C40, 60G60.

Key words and phrases. nonparametric regression; graphs; multidimensional wavelets; spatial lattice processes; strong spatial mixing; Bernstein inequality; road networks.

The author gratefully acknowledges the financial support of the Fraunhofer ITWM which is part of the Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.

Definition 1.1 (Random field). Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let V be an index set and let (S_v, \mathfrak{S}_v) be a measurable space for $v \in V$. Let $Z := \{Z(v) : v \in V\}$ be a set of random variables on $(\Omega, \mathcal{A}, \mathbb{P})$ such that each $Z(v)$ takes values in (S_v, \mathfrak{S}_v) . Then, the collection Z is called a random field.

Definition 1.2 (Homogeneous random field). Let $(\Gamma, +)$ be a group. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space endowed with the random field $\{Z(s) : s \in \Gamma\}$ where each $Z(s)$ takes values in the same state space (S, \mathfrak{S}) . The random field is called homogeneous or stationary if for each $n \in \mathbb{N}_+$ and for all points $s_1, \dots, s_n \in \Gamma$ and each translation $t \in \Gamma$ the joint probability distribution of the collection $\{Z(s_1 + t), \dots, Z(s_n + t)\}$ is identical with the joint probability distribution of $\{Z(s_1), \dots, Z(s_n)\}$.

Denote by $\|\cdot\|_\infty$ the maximum norm on \mathbb{R}^N and by d_∞ the corresponding metric which is extended to subsets I, J of \mathbb{R}^N via $d_\infty(I, J) := \inf\{d_\infty(s, t) : s \in I, t \in J\}$. Furthermore, write $s \leq t$ for $s, t \in \mathbb{R}^N$ if and only if for each $1 \leq k \leq N$ the single coordinates satisfy $s_k \leq t_k$.

Definition 1.3 (Strong spatial mixing). Let $\{Z(s) : s \in \Gamma\}$ be a random field for $\Gamma \subseteq \mathbb{Z}^N$, $N \in \mathbb{N}_+$. Denote for a subset I of Γ by $\mathcal{F}(I) = \sigma(Z(s) : s \in I)$ be the σ -algebra generated by the $Z(s)$ in I . Define for $n \in \mathbb{N}_+$ the α -mixing coefficient by

$$\alpha(n) := \sup_{\substack{I, J \subseteq \Gamma, \\ d_\infty(I, J) \geq n}} \sup_{\substack{A \in \mathcal{F}(I), \\ B \in \mathcal{F}(J)}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|$$

The random field is strong spatial mixing if $\alpha(n) \rightarrow 0$ as $n \rightarrow \infty$.

We shall work on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ which is equipped with a strong spatial mixing random field $Z := \{Z(s) : s \in I\}$ such that each $Z(s)$ takes values in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ for a subset $I \subseteq \mathbb{Z}^N$ ($N \geq 1$) where $I^+ := I \cap \mathbb{N}_+^N$ is infinite. We precise this with the following condition:

Condition 1.4 (Regularity condition for random fields). Let $I \subseteq \mathbb{Z}^N$, $N \in \mathbb{N}_+$, be such that $I^+ := I \cap \mathbb{N}_+^N$ is infinite. Let $Z = \{Z(s) : s \in I^+\}$ be a random field such that each $Z(s)$ takes values in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and has equal marginal distributions, i.e., $\mathcal{L}_{Z(s)} = \mathcal{L}_{Z(t)}$. Furthermore, Z is strong mixing with exponentially decreasing mixing coefficients: there are $c_0, c_1 \in \mathbb{R}_+$ such that $\alpha(k) \leq c_0 \exp(-c_1 k)$ for all $k \in \mathbb{N}_+$.

Let $(n(k) : k \in \mathbb{N}_+) \subseteq \mathbb{N}_+^N$ be an increasing sequence in that $e_N \leq n(k) \leq n(k+1)$. The sequence fulfills both

$$\liminf_{k \rightarrow \infty} \min_{1 \leq i \leq N} n_i(k) \geq e^2 \text{ and } \liminf_{k \rightarrow \infty} \max_{1 \leq i \leq N} n_i(k) = \infty \text{ as } k \rightarrow \infty.$$

Define the increasing sequence of index sets by $I_{n(k)} := \{s \in I^+ : s \leq n(k)\} \subseteq \mathbb{N}_+^N$. $I_{n(1)}$ contains the element $e_N = (1, \dots, 1)^T$. The set I together with the sets $I_{n(k)}$ satisfies the growth condition, $|I_{n(k)}| \geq C \left(\prod_{i=1}^N n_i(k) \right)^\rho$, for $\frac{N}{N+1} < \rho \leq 1$ and some $0 < C < \infty$.

With this condition, we ensure on the one hand that there are sufficiently many data points selected in the sampling process by requiring that the running maximum converges to infinity (the condition on the running minimum is technical). And on the other hand we do not rule out the possibility to omit certain points from the lattice, e.g., by choosing $\rho < 1$. This can prove convenient in theoretical applications where the data structure is an infinite graph which differs from the regular lattice by a "certain amount of wholes". Next, we define the covering number for function classes.

Definition 1.5 (ε -covering number). Let $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ be endowed with a probability measure ν and let \mathcal{G} be a set of real valued Borel functions on \mathbb{R}^d and let $\varepsilon > 0$. Every finite collection g_1, \dots, g_N of Borel functions on \mathbb{R}^d is called an ε -cover of \mathcal{G} w.r.t. $\|\cdot\|_{L^p(\nu)}$ of size N if for each $g \in \mathcal{G}$ there is a j , $1 \leq j \leq N$, such that $\|g - g_j\|_{L^p(\nu)} < \varepsilon$. The ε -covering number of \mathcal{G} w.r.t. $\|\cdot\|_{L^p(\nu)}$ is defined as

$$N(\varepsilon, \mathcal{G}, \|\cdot\|_{L^p(\nu)}) := \inf \left\{ N \in \mathbb{N} : \exists \varepsilon\text{-cover of } \mathcal{G} \text{ w.r.t. } \|\cdot\|_{L^p(\nu)} \text{ of size } N \right\}.$$

The covering number is monotone: $N(\varepsilon_2, \mathcal{G}, \|\cdot\|_{L^p(\nu)}) \leq N(\varepsilon_1, \mathcal{G}, \|\cdot\|_{L^p(\nu)})$ if $\varepsilon_1 \leq \varepsilon_2$.

The covering number can be bounded uniformly over all probability measures for a class of bounded functions under mild regularity conditions. Thus, the following covering condition is appropriate for many function classes \mathcal{G} .

Condition 1.6 (Covering condition). \mathcal{G} is a class of uniformly bounded, measurable functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\|f\|_\infty \leq B < \infty$ and for all $\varepsilon > 0$ and all $N \geq 1$ the following is true:

For any choice $z_1, \dots, z_N \in \mathbb{R}^d$ the ε -covering number of \mathcal{G} w.r.t. the L^1 -norm of the discrete measure with point masses $\frac{1}{N}$ in z_1, \dots, z_N is bounded by a deterministic function depending only on ε and \mathcal{G} , which we shall denote by $H_{\mathcal{G}}(\varepsilon)$, i.e., $N(\varepsilon, \mathcal{G}, \frac{1}{N} \sum_{k=1}^N \delta_{z_k}) \leq H_{\mathcal{G}}(\varepsilon)$.

We are now prepared to introduce the nonparametric regression model: let there be given the random field (X, Y) which satisfies Condition 1.4. The random variables $X(s)$ are \mathbb{R}^d -valued and have equal marginal distributions denoted by the probability measure μ_X on \mathbb{R}^d . The $Y(s)$ are \mathbb{R} -valued and satisfy for each $s \in I$ the relation

$$Y(s) = m(X(s)) + \varsigma(X(s)) \varepsilon(s), \quad (1.1)$$

where $m, \varsigma : \mathbb{R}^d \rightarrow \mathbb{R}$ are functions in $L^2(\mu_X)$ and the error terms $\varepsilon(s) \sim (0, 1)$ are independent from X and have identical marginal distributions but may be dependent among each other such that the strong mixing property remains valid. Note that we do not require any specific distribution of the error terms, e.g., a Gaussian distribution. In addition, let $\mathcal{F}_k \subseteq L^2(\mu_X)$ for $k \in \mathbb{N}_+$ be a deterministic sequence of increasing function classes whose union is dense in $L^2(\mu_X)$. We define for $k \in \mathbb{N}_+$ the least-squares minimizer

$$m_k := \arg \min_{f \in \mathcal{F}_k} |I_{n(k)}|^{-1} \sum_{s \in I_{n(k)}} (Y(s) - f(X(s)))^2, \quad (1.2)$$

where $I_{n(k)} \subseteq I_{n(k+1)} \subseteq I^+$ is increasing such that $\cup_{k \in \mathbb{N}} I_{n(k)} = I^+$. We shall choose the \mathcal{F}_k in applications as finite dimensional linear spaces induced by real valued functions $f_1, \dots, f_{K_k} : \mathbb{R}^d \rightarrow \mathbb{R}$, i.e.,

$$\mathcal{F}_k = \left\{ \sum_{j=1}^{K_k} a_j f_j : a_j \in \mathbb{R}, j = 1, \dots, K_k \right\}. \quad (1.3)$$

But we formulate the subsequent statements for general deterministic classes of functions \mathcal{F}_k . Using linear spaces as \mathcal{F}_k has the computational advantage that the minimization is an unrestricted ordinary least-squares problem on the domain of the parameters without an additional penalizing term. However, in order to render the estimator robust against deviations in the data, we consider the truncated estimator which is defined for a real-valued sequence $\{\beta_k : k \in \mathbb{N}_+\}$ which converges to infinity as

$$\hat{m}_k := T_{\beta_k} m_k \quad (1.4)$$

where for $L > 0$ the truncation operator is $T_L y := \max(\min(y, L), -L)$. We are now prepared to state results on the consistency of the truncated least-squares estimator \hat{m}_k from equations (1.1), (1.2) and (1.4):

Theorem 1.7 (Consistency of truncated least-squares on N -dimensional lattices). *Let the random field (X, Y) from equation (1.1) satisfy Condition 1.4. Let the $Y(s)$ be square integrable and denote by μ_X the marginal law of the $X(s)$ on \mathbb{R}^d . Let the \mathcal{F}_k be increasing function classes $f : \mathbb{R}^d \rightarrow \mathbb{R}$ whose union is dense in $L^2(\mu_X)$. Let the map*

$$\Omega \ni \omega \mapsto \sup_{f \in T_{\beta_k} \mathcal{F}_k} \left| \frac{1}{|I_{n(k)}|} \sum_{s \in I_{n(k)}} |f(X(s, \omega)) - T_L Y(s, \omega)|^2 - \mathbb{E} \left[|f(X(e_N)) - T_L Y(e_N)|^2 \right] \right| \quad (1.5)$$

be measurable and let Condition 1.6 prevail. Put for short hand

$$\kappa_k(\varepsilon, \beta_k) := \log H_{T_{\beta_k} \mathcal{F}_k} \left(\frac{\varepsilon}{128 \beta_k} \right).$$

Assume that both $\beta_k \rightarrow \infty$ and that $\kappa_k(\varepsilon, \beta_k) \rightarrow \infty$ as $k \rightarrow \infty$. If

$$\beta_k^2 \kappa_k(\varepsilon, \beta_k) \left(\prod_{i=1}^N \log n_i(k) \right) / \left(\prod_{i=1}^N n_i(k) \right)^{\rho - N/(N+1)} \rightarrow 0 \text{ as } k \rightarrow \infty,$$

the sequence of estimators $\{\hat{m}_k : k \in \mathbb{N}_+\}$ is weakly universally consistent, i.e.,

$$\lim_{k \rightarrow \infty} \mathbb{E} \left[\int_{\mathbb{R}^d} (\hat{m}_k - m)^2 d\mu_X \right] = 0.$$

If additionally, Y is ergodic in the sense that

$$\frac{1}{|I_{n(k)}|} \sum_{s \in I_{n(k)}} |Y(s) - T_L Y(s)|^2 \rightarrow \mathbb{E} \left[|Y(s) - T_L Y(s)|^2 \right] \quad \text{a.s. for all } L > 0$$

and if there exists a $\delta > 0$ such that

$$\left\{ \beta_k^2 \left(\prod_{i=1}^N \log n_i(k) \right) (\log k)^{1+\delta} \right\} / \left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)} \rightarrow 0 \text{ as } k \rightarrow \infty,$$

then $\{\hat{m}_k : k \in \mathbb{N}_+\}$ is strongly universally consistent, i.e., $\lim_{k \rightarrow \infty} \int_{\mathbb{R}^d} (\hat{m}_k - m)^2 d\mu_X = 0$ on a measurable set $\Omega_0 \in \mathcal{A}$ with $\mathbb{P}(\Omega_0) = 1$.

In case where the conditions of Theorems A.1 and Theorem B.2 are satisfied, we can formulate a useful corollary:

Corollary 1.8. *Let (X, Y) be a stationary random field on a full lattice $I = \mathbb{Z}^N$, $N \in \mathbb{N}_+$ such that Condition 1.4 is fulfilled. Let $(K_k : k \in \mathbb{N}_+) \subseteq \mathbb{N}_+$ be a sequence converging to infinity. Let \mathcal{F}_k be the linear span of continuous, linear independent functions f_1, \dots, f_{K_k} , as in (1.3) such that $\cup_{k \in \mathbb{N}_+} \mathcal{F}_k$ is dense in $L^2(\mu_X)$. Let the index sets be defined by the canonical sequence $n(k) := k \cdot e_N \in \mathbb{N}_+^N$. The estimator is weakly universally consistent if*

$$K_k \beta_k^2 \log \beta_k (\log k)^N / k^{N/(N+1)} \rightarrow 0 \text{ as } k \rightarrow \infty.$$

The estimator is strongly universally consistent if additionally

$$\beta_k^2 (\log k)^{N+2} / k^{N/(N+1)} \rightarrow 0 \text{ as } k \rightarrow \infty.$$

The next results concern the rate of convergence of the truncated least-squares estimator \hat{m}_k from equations (1.1), (1.2) and (1.4). In this analysis, we encounter an empirical error which depends on the chosen $\omega \in \Omega$ and an approximation error which relates the function m to its projection onto the function classes \mathcal{F}_k .

As we did not rule out dependence among the error terms $\varepsilon(s)$, the conditional covariance between two distinct observations $Y(s)$ and $Y(t)$ is in general not zero. Thus, we need a condition on the conditional covariance matrix of the observations $Y(s)$ which we denote by $\text{Cov}(Y(I_{n(k)}) | X(I_{n(k)}))$. Mark that in the special case where the error terms are uncorrelated, $\text{Cov}(Y(I_{n(k)}) | X(I_{n(k)}))$ is a diagonal matrix and it is sufficient to impose a restriction on the conditional variances. We state the second main theorem:

Theorem 1.9 (Rate of convergence). *Let (X, Y) be the random field from equation (1.1) which satisfies Condition 1.4 such that $\|m\|_\infty \leq L$. Let the conditional variance fulfill $\sup_{x \in \mathbb{R}^d} \text{Var}(Y(e_N) | X(e_N) = x) < \infty$. If the error terms $\varepsilon(s)$ are not independent, let there exist a $\gamma > 0$ such that $\mathbb{E} [|\varepsilon(e_N)|^{2+\gamma}] < \infty$. Let the function classes \mathcal{F}_k be defined by equation (1.3) as linear spaces. Let*

$$K_k \left(\prod_{i=1}^N \log n_i(k) \right)^3 / \left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)} \rightarrow 0 \text{ as } k \rightarrow \infty.$$

Then there is a universal constant $0 < C < \infty$ such that for all $k \in \mathbb{N}_+$

$$\mathbb{E} \left[\int_{\mathbb{R}^d} |\hat{m}_k - m|^2 d\mu_X \right] \leq 8 \inf_{f \in \mathcal{F}_k} \int_{\mathbb{R}^d} |f - m|^2 d\mu_X + C \frac{K_k \left(\prod_{i=1}^N \log n_i(k) \right)^3}{\left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)}}.$$

For the case of an i.i.d. sample Györfi et al. [2002] find that under the same assumptions the estimation error can be bounded by $K_k(\log k + 1)/k$ times a constant and for a sample of size k . In the following, we shall investigate the practical problem

$$\arg \min_{f \in \mathcal{F}_k} \frac{1}{|I_{n(k)}|} \sum_{s \in I_{n(k)}} (f(X(s)) - Y(s))^2 = \arg \min_{a \in \mathbb{R}^{K_k}} \frac{1}{|I_{n(k)}|} \sum_{s \in I_{n(k)}} \left(\sum_{i=1}^{K_k} a_i f_i(X(s)) - Y(s) \right)^2$$

where the function classes are given as linear spaces as in equation (1.3) and which leads to a linear least-squares problem $\min_{a \in \mathbb{R}^v} \|Za - y\|_2^2$. The matrix Z contains in the $v = K_k$ columns the basis functions f_i

evaluated at the data vector $(X(s) : s \in I_{n(k)}) \in \mathbb{R}^{|I_{n(k)}| \times d}$. This means, let an enumeration $\iota : \mathbb{N}_+ \rightarrow \mathbb{N}_+^d$ of the spatial coordinates of \mathbb{N}_+^d as a subset of \mathbb{Z}^d be given, then

$$Z = \begin{pmatrix} f_1(X(\iota(1))) & \dots & f_v(X(\iota(1))) \\ \vdots & \ddots & \vdots \\ f_1(X(\iota(|I_{n(k)}|))) & \dots & f_v(X(\iota(|I_{n(k)}|))) \end{pmatrix}$$

Since in general this matrix Z might not have full rank, the usual linear regression routine which requires no multicollinearity, can break down. We remedy this problem with the principal component regression and the singular value decomposition. Let $U\Sigma V^T \in \mathbb{R}^{m \times n}$ be a singular value decomposition of the real valued matrix Z where $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{m \times n}$ is a (rectangular) diagonal matrix of the type $\sigma_1 \geq \dots \geq \sigma_r > 0 = \sigma_{r+1} = \dots = \sigma_{\min(m,n)}$. We make the convention to write $z = V^T \cdot a$ and $U = (u_1, \dots, u_m)$ where u_i are the column vectors of U . We get using the fact that orthogonal matrices preserve lengths and angles

$$\begin{aligned} \|Za - y\|_2^2 &= \|U\Sigma V^T a - y\|_2^2 = \|U(\Sigma V^T a - U^T y)\|_2^2 = \|\Sigma z - U^T y\|_2^2 \\ &= \sum_{i=1}^r (\sigma_i z_i - (u_i)^T y)^2 + \sum_{i=r+1}^m ((u_i)^T y)^2. \end{aligned}$$

Hence, all solutions to the linear regression problem are given by $a = Vz$ with $z_i = (u_i)^T y / \sigma_i$ for $i = 1, \dots, r$ and z_i arbitrary, for $i = r+1, \dots, n$. In particular, upon choosing $z_i = 0$ for $i = r+1, \dots, n$, we can write the solution associated with this choice as $a^* = \sum_{i=1}^r (u_i)^T y v_i / \sigma_i$. It is straightforward to apply this technique to the nonparametric estimator \hat{m}_k given in equations (1.2) and (1.4).

2. LINEAR WAVELET REGRESSION ON STRONG SPATIAL MIXING DATA

Since we use wavelets for numerical applications of the above developed asymptotic theory, we give a review on important concepts of wavelets in d dimensions, the definitions are taken from the monograph of Benedetto [1993].

Definition 2.1 (Multiresolution Analysis). Let $\Gamma \subseteq \mathbb{R}^d$ be a lattice, this is a discrete subgroup given by $(\Gamma, +) = (\{\sum_{i=1}^d a_i v_i : a_i \in \mathbb{Z}\}, +)$ for certain $v_i \in \mathbb{R}^d$ ($i = 1, \dots, d$). Furthermore, let $M \in \mathbb{R}^{d \times d}$ be a matrix which preserves the lattice Γ , i.e., $M\Gamma \subseteq \Gamma$ and which is strictly expanding, i.e., all eigenvalues λ of M satisfy $|\lambda| > 1$. Denote for such a matrix M the absolute value of its determinant by $|M|$. A multiresolution analysis (MRA) of $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d)$, $d \in \mathbb{N}_+$, with a scaling function $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is an increasing sequence of subspaces of $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d)$ given by $\dots \subseteq U_{-1} \subseteq U_0 \subseteq U_1 \subseteq \dots$ such that the following four conditions are satisfied

- (1) (Denseness) $\bigcup_{j \in \mathbb{Z}} U_j$ is dense in $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d)$,
- (2) (Separation) $\bigcap_{j \in \mathbb{Z}} U_j = \{0\}$,
- (3) (Scaling) $f \in U_j$ if and only if $f(M^{-j} \cdot) \in U_0$,
- (4) (Orthonormality) $\{\Phi(\cdot - \gamma) : \gamma \in \Gamma\}$ is an orthonormal basis of U_0 .

It is straightforward to show that given an MRA with corresponding scaling function Φ there is a sequence $(a_0(\gamma) : \gamma \in \Gamma) \subseteq \mathbb{R}$ such that $\Phi \equiv \sum_{\gamma \in \Gamma} a_0(\gamma) \Phi(M \cdot - \gamma)$ and the coefficients $a_0(\gamma)$ fulfill the equations $a_0(\gamma) = |M| \int_{\mathbb{R}^d} \Phi(x) \Phi(Mx - \gamma) dx$ and $\sum_{\gamma \in \Gamma} |a_0(\gamma)|^2 = |M| = \sum_{\gamma \in \Gamma} a_0(\gamma)$.

In the following, we write $L^2(\lambda^d)$ for $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d)$. The relationship between an MRA and an orthonormal basis of $L^2(\lambda^d)$ is summarized in the next theorem. We have

Theorem 2.2 (Benedetto [1993]). Suppose Φ generates a multiresolution analysis and the $a_k(\gamma)$ satisfy for all $0 \leq j, k \leq |M| - 1$ and $\gamma \in \Gamma$ the equations

$$\sum_{\gamma' \in \Gamma} a_j(\gamma') a_k(M\gamma + \gamma') = |M| \delta(j, k) \delta(\gamma, 0) \quad \text{and} \quad \sum_{\gamma \in \Gamma} a_0(\gamma) = |M|.$$

Furthermore, let for $k = 1, \dots, |M| - 1$ the functions Ψ_k be given by $\Psi_k := \sum_{\gamma \in \Gamma} a_k(\gamma) \Phi(M \cdot - \gamma)$. Then the set of functions $\{|M|^{j/2} \Psi_k(M^j \cdot - \gamma) : j \in \mathbb{Z}, k = 1, \dots, |M| - 1, \gamma \in \Gamma\}$ form an orthonormal basis of $L^2(\lambda^d)$:

$$L^2(\lambda^d) = U_0 \oplus \left(\bigoplus_{j \in \mathbb{N}} W_j \right) = \bigoplus_{j \in \mathbb{Z}} W_j,$$

$$\text{where } W_j := \langle |M|^{j/2} \Psi_k(M^j \cdot - \gamma) : k = 1, \dots, |M| - 1, \gamma \in \Gamma \rangle.$$

We sketch in a short example how to construct a d -dimensional MRA given that one has a father and a mother wavelet on the real line.

Example 2.3 (Isotropic d -dimensional MRA from one-dimensional MRA via tensor products). Let $d \in \mathbb{N}_+$ and let φ be a scaling function on the real line \mathbb{R} together with the mother wavelet ψ which fulfill the equation

$$\varphi \equiv \sqrt{2} \sum_{k \in \mathbb{Z}} h_k \varphi(2 \cdot -k) \text{ and } \psi \equiv \sqrt{2} \sum_{k \in \mathbb{Z}} g_k \varphi(2 \cdot -k),$$

for real sequences $(h_k : k \in \mathbb{Z})$ and $(g_k : k \in \mathbb{Z})$. Let φ generate an MRA of $L^2(\lambda)$ with the corresponding spaces U'_j , $j \in \mathbb{Z}$. The d -dimensional wavelets are derived as follows: put $\Gamma := \mathbb{Z}^d$ and define the diagonal matrix M by $M := 2 \operatorname{diag}(1, \dots, 1)$. Furthermore, set $\xi_0 := \varphi$ and $\xi_1 := \psi$. Denote the mother wavelets as pure tensors by $\Psi_k := \xi_{k_1} \otimes \dots \otimes \xi_{k_d}$ for $k \in \{0, 1\}^d \setminus 0$. The scaling function is given as $\Phi := \Psi_0 := \otimes_{i=1}^d \varphi$.

Then, as demonstrated in Appendix, Φ and the linear spaces $U_j := \otimes_{i=1}^d U'_j$ form an MRA of $L^2(\lambda^d)$ and the functions Ψ_k , $k \neq 0$, generate an orthonormal basis in that

$$L^2(\lambda^d) = U_0 \oplus \left(\oplus_{j \in \mathbb{N}} W_j \right) = \oplus_{j \in \mathbb{Z}} W_j$$

where $W_j = \left\langle |M|^{j/2} \Psi_k (M^j \cdot -\gamma) : \gamma \in \mathbb{Z}^d, k \in \{0, 1\}^d \setminus 0 \right\rangle$.

In the sequel, we bridge the gap between nonparametric regression and wavelet theory. From Theorem 1.7 we infer that the function spaces \mathcal{F}_k need to densely approximate $L^2(\mu_X)$ for any probability measure μ_X . The next theorem states that wavelets fulfill this condition.

Theorem 2.4 (Wavelets are dense in $L^p(\mu)$ for isotropic MRA). *Let there be given an isotropic MRA on \mathbb{R}^d , $d \geq 1$ with corresponding scaling function Φ constructed as in Example 2.3 from a compactly supported real scaling function φ . Let μ be a probability measure on $\mathcal{B}(\mathbb{R}^d)$ and let $1 \leq p < \infty$, then $\cup_{j \in \mathbb{Z}} U_j$ is dense in $L^p(\mu)$.*

We intend to estimate a random field (X, Y) which satisfies Condition 1.4 with a nonparametric wavelet estimator as follows: let an MRA of $L^2(\lambda^d)$ with compactly supported wavelets be given. Put for short $\Phi_{j,\gamma} := |M|^{j/2} \Phi(M^j \cdot -\gamma)$ where Φ is the corresponding scaling function and M is an expanding matrix ($\gamma \in \mathbb{Z}^d$ and $j \in \mathbb{Z}$). Define for two increasing sequences $(w_k : k \in \mathbb{N}) \subseteq \mathbb{Z}$ and $(j(k) : k \in \mathbb{N}) \subseteq \mathbb{Z}$ with $\lim_{k \rightarrow \infty} w_k = \lim_{k \rightarrow \infty} j(k) = \infty$ the set $K_k := \{\gamma \in \mathbb{Z}^d : \|\gamma\|_\infty \leq w_k\} \subseteq \mathbb{Z}^d$. Furthermore, define for $k \in \mathbb{N}_+$ the linear space

$$\mathcal{F}_k := \left\{ \sum_{\gamma \in K_k} a_\gamma \Phi_{j(k),\gamma} : a_\gamma \in \mathbb{R} \right\} \subseteq U_{j(k)}. \quad (2.1)$$

With the help of Corollary 1.8 and Theorem 1.9 we can formulate two theorems. Therefore, let M be a diagonalizable matrix, $M = S^{-1}DS$ where D is a diagonal matrix containing the eigenvalues of M . Denote by $\lambda_{\max} := \max\{|\lambda_i| : i = 1, \dots, d\}$ the maximum of the absolute values of the eigenvalues. We define the 2-norm of a square matrix $A = (a_{i,j})_{1 \leq i,j \leq d} \in \mathbb{R}^{d \times d}$ as $\|A\|_2 = \max_{x: \|x\|_2=1} \|Ax\|_2$.

Theorem 2.5 (Consistency of wavelet based linear regression). *Let the function $m \in L^2(\mu_X)$. Let the random field (X, Y) be defined on a full N -dimensional lattice and let the wavelet basis be dense in $L^2(\mu_X)$. Let $\beta_k := c \log k$ for some constant $c \in \mathbb{R}_+$. The wavelet based estimator \hat{m}_k from equations (1.1), (1.2), (1.4) and 2.1 is weakly universally consistent if*

$$\lim_{k \rightarrow \infty} (\lambda_{\max})^{j(k)} / w_k = 0 \text{ and}$$

$$\lim_{k \rightarrow \infty} w_k^d (\log k)^2 \log \log k \prod_{i=1}^N \log n_i(k) \left/ \left(\prod_{i=1}^N n_i(k) \right)^{1/(N+1)} \right. = 0.$$

Furthermore, let (X, Y) be defined on entire \mathbb{Z}^N . The estimator is strongly universally consistent if additionally (X, Y) is stationary and if

$$\lim_{k \rightarrow \infty} (\log k)^4 \prod_{i=1}^N \log n_i(k) \left/ \left(\prod_{i=1}^N n_i(k) \right)^{1/(N+1)} \right. = 0.$$

Theorem 2.6 (Rate of convergence of wavelet based linear regression). *Under the conditions of Theorem 2.5. If $\|m\|_\infty < \infty$ and under the additional assumptions of Theorem 1.9 there is a constant C which does not depend*

on k such that the rate of convergence of the estimator is at least

$$\mathbb{E} \left[\int_{\mathbb{R}^d} (\hat{m}_k - m)^2 d\mu_X \right] \leq C w_k^d \left(\prod_{i=1}^N \log n_i(k) \right)^3 / \left(\prod_{i=1}^N n_i(k) \right)^{1/(N+1)} \\ + 8 \inf_{f \in \mathcal{F}_k} \int_{\mathbb{R}^d} (f - m)^2 d\mu_X.$$

We give a short application for an isotropic Haar basis in d -dimensions in the case where the regression function m is (A, r) -Hölder continuous, that is for all $x, y \in \text{dom}(m)$

$$|m(x) - m(y)| \leq A \|x - y\|_\infty^r \text{ for } A \in \mathbb{R}_+ \text{ and } r \in (0, 1].$$

Corollary 2.7 (Rate of convergence for Hölderian functions). *Let the conditions of Theorem 2.6 be fulfilled and let the $X(s)$ satisfy $\mathbb{P}(\|X(s)\|_\infty > t) \in \mathcal{O}(t^{-2})$. Let the conditional mean function m be (A, r) -Hölder continuous. Define the resolution index as*

$$j(k) := \left\lfloor \frac{1/\log 2}{d+2r} \log R(k) - \frac{d/\log 2}{d+2r} \log h(k) \right\rfloor \text{ where } R(k) := \frac{\left(\prod_{i=1}^N n_i(k) \right)^{1/(N+1)}}{\left(\prod_{i=1}^N \log n_i(k) \right)^3}$$

is the cross convergence rate from Theorem 2.6 and h is a positive function with $\lim_{k \rightarrow \infty} h(k) = \infty$ and $\log h(k) \in o(\log R(k))$. Define the window as $w_k := 2^{j(k)} h(k)$. Then the MISE satisfies

$$\mathbb{E} \left[\int_{\mathbb{R}^d} |\hat{m}_k - m|^2 d\mu \right] \in \mathcal{O} \left(R(k)^{-2r/(d+2r)} h(k)^{2rd/(d+2r)} \right). \quad (2.2)$$

In particular, for the canonical index sets $I_{n(k)}$ defined with $n(k) := k e_N$ and a resolution index as

$$j(k) := \left\lfloor \frac{N/(N+1)}{\log 2(d+2r)} \log k - \frac{1/\log 2}{d+2r} \{3N \log \log k + d \log h(k)\} \right\rfloor$$

the MISE fulfills

$$\mathbb{E} \left[\int_{\mathbb{R}^d} |\hat{m}_k - m|^2 d\mu \right] \in \mathcal{O} \left(k^{-(N/(N+1) 2r/(d+2r))} (\log k)^{3N 2r/(d+2r)} h(k)^{2rd/(d+2r)} \right).$$

Proof. Note that by construction $\|M\|_2^{j(k)} / w_k \rightarrow 0$ and that the estimation error is contained in the right-hand side of (2.2). It remains to compute the approximation error: there is a function $f \in \mathcal{F}_k$ which is piecewise constant on dyadic d -dimensional cubes of edge length 2^{-j} with values

$$f(x) = m \left((a_1, \dots, a_d) / 2^j \right) \text{ for } x \in \left[(a_1, \dots, a_d) / 2^j, ((a_1, \dots, a_d) + e_N) / 2^j \right),$$

where $a_i \in \mathbb{Z}$ for $i = 1, \dots, d$. For this f we have

$$\int_{\mathbb{R}^d} |f - m|^2 d\mu_X \leq \sup_{\text{dom } f} |f - m|^2 + \int_{\mathbb{R}^d \setminus \text{dom } f} m^2 d\mu_X.$$

The first term is at most $A^2 2^{-2rj(k)}$ by construction and obviously attains the stated rate. The second term behaves as $\mathbb{P}(\|X(e_N)\|_\infty > w_k/2) \in \mathcal{O}(w_k^{-2})$ which is again in the right-hand side of (2.2). \square

For the particular case that the $X(s)$ are bounded, we obtain in the same way as in Corollary 2.7 a slightly better rate because in this case it suffices that the effective window size $w_k / \|M\|_2^{j(k)}$ remains constant and h can be chosen as a constant. With canonical index sets the convergence rate of the L^2 -error reduces to

$$\mathbb{E} \left[\int_{\mathbb{R}^d} |\hat{m}_k - m|^2 d\mu \right] \in \mathcal{O} \left(k^{-(N/(N+1) 2r/(d+2r))} (\log k)^{3N 2r/(d+2r)} \right).$$

3. EXAMPLES OF APPLICATION

3.1. Simulation concepts for Markov random fields. This subsection introduces an algorithm to simulate (Markov) random fields that are defined on arbitrary graphs $G = (V, E)$ with a finite set of nodes V . The main idea dates back at least to Kaiser et al. [2012] and is based on the concept of *concliques* which has the advantage that simulations can be performed faster when compared to the Gibbs sampler; an introduction to Gibbs sampling offers Brémaud [1999]. We start with a definition

Definition 3.1 (Concliques, cf. Kaiser et al. [2012]). Let $G = (V, E)$ be an undirected graph with a countable set of nodes V and let $C \subseteq V$. If for all pairs of nodes $(v, w) \in C \times C$ satisfy $\{v, w\} \notin E$, the set C is called a conclique. A collection C_1, \dots, C_n of concliques that partition V is called a conclique cover; the collection is a minimal conclique cover if it contains the smallest number of concliques needed to partition V .

Definition 3.2 (Full conditional distribution). Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let (S, \mathfrak{S}) be a state space. Let $Y = \{Y(s) : s \in I\}$ be a collection of S -valued random variables. Then we call the family $\{\mathbb{P}(Y(s) \in \cdot \mid Y(t), t \in I \setminus \{s\})\}$ a full conditional distribution of Y .

Let G be a finite graph whose nodes are partitioned into a conclique cover C_1, \dots, C_n . Denote by $Ne(v)$ the neighbors of v in G for $v \in V$. Let $Y = (Y(v) : v \in V)$ be a Markov random field on G which takes values in (S, \mathfrak{S}) with a full conditional distribution $\{F_v(Y(v) \in A \mid Y(w), w \in Ne(v)) : v \in V\}$ and an initial distribution μ_0 . Note that the joint conditional distribution of a conclique $Y(C_i)$ given its neighbors which are contained in $Y(C_1), \dots, Y(C_{i-1}), Y(C_{i+1}), \dots, Y(C_n)$ factorizes as the product of the single conditional distributions due to the Markov property. This entails that we can – under mild regularity conditions – simulate the stationary distribution of the MRF with a Markov chain using the following algorithm:

Algorithm 3.3 (Simulation of random fields with concliques, Kaiser et al. [2012]). Simulate the starting values according to an initial distribution μ_0 and obtain the vector of $Y^{(0)} = (Y^{(0)}(C_1), \dots, Y^{(0)}(C_n))$.

In the next step, given a vector $Y^{(k)} = (Y^{(k)}(C_1), \dots, Y^{(k)}(C_n))$, simulate the concliques $Y^{(k+1)}(C_i)$ given the $(k+1)$ -st simulation of the neighbors in $Y^{(k+1)}(C_1), \dots, Y^{(k+1)}(C_{i-1})$ and k -th simulation of the neighbors in $Y^{(k)}(C_{i+1}), \dots, Y^{(k)}(C_n)$ with the specified full conditional distribution for $i = 1, \dots, n$. Repeat this step, until the maximum iteration number for the index k is reached.

In the sequel, we formally describe the Markov kernel of the Markov chain $\{Y^{(k)} : k \in \mathbb{N}\}$ for the case where the full conditional distribution is specified in terms of conditional densities. We assume that (S, \mathfrak{S}) is equipped with a σ -finite measure ν such that the distribution of Y is absolutely continuous with respect to ν , i.e., $\mathbb{P}_Y \ll \nu$ with a density f . We write for convenience $C_{-I} := \cup_{i \notin I} C_i$ for the conclique cover C_1, \dots, C_n , for $I \subseteq \{1, \dots, n\}$. Furthermore, let an enumeration within each conclique i be given by $C_i = \{(i, 1), \dots, (i, l_i)\}$. Denote the conditional density of the node (i, s) given its neighbors by $f_{(i,s) \mid Ne(i,s)}$, then the transition kernel which captures the evolution of $Y(C_i)$ given $Y(C_{-i})$ is given by

$$\mathbb{M}_i : S^{|C_{-i}|} \times \mathfrak{S}^{|C_i|} \rightarrow [0, 1],$$

$$(y(C_{-i}), B) \mapsto \int_B \prod_{s=1}^{l_i} f_{(i,s) \mid Ne(i,s)}(y(i, s) \mid y(Ne(i, s))) \nu^{\otimes C_i}(dy(C_i)). \quad (3.1)$$

With the help of (3.1) the Markov kernel for the entire chain $\{Y^{(k)} : k \in \mathbb{N}\}$ can be written as

$$\mathbb{M} : S^{|V|} \times \mathfrak{S}^{|V|} \rightarrow [0, 1],$$

$$(y, B) \mapsto \int_{S^{|C_1|}} M_1(y(C_{-1}), dx(C_1)) \int_{S^{|C_2|}} M_2((x(C_1), y(C_{-1,2})), dx(C_2)) \dots$$

$$\dots \int_{S^{|C_i|}} M_i((x(C_1), \dots, x(C_{i-1}), y(C_{i+1}), \dots, y(C_n)), dx(C_n)) \dots$$

$$\dots \int_{S^{|C_n|}} M_n((x(C_{-n})), dx(C_n)) 1_B(x). \quad (3.2)$$

We are able to prove with these definitions

Theorem 3.4. *Let the density f be strictly positive on $S^{\times |V|}$ such that the conditional densities $f_{C(i,s) \mid Ne(i,s)}$ furnish a full conditional distribution, then the distribution of Y , \mathbb{P}_Y , is an invariant probability measure of the Markov chain given by equations (3.1) and (3.2) in the sense that $\mathbb{P}_Y \mathbb{M} \equiv \mathbb{P}_Y$. That is \mathbb{M} is positive.*

It remains to prove the accuracy of the simulation approach of the homogeneous Markov chain simulated from a Markov random field as proposed in Algorithm 3.3 and equations (3.1) and (3.2) in the case that $(S, \mathfrak{S}) \subseteq (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. This means, we ask whether the chain is ergodic in the sense that $\lim_{n \rightarrow \infty} \|\nu_0 \mathbb{M}^n - \mathbb{P}_Y\|_{TV} = 0$ in the total variation norm for the positive Markov kernel \mathbb{M} with invariant probability measure \mathbb{P}_Y and for all distributions ν_0 on $\mathfrak{S}^{|V|}$.

Theorem 3.5. *Let the Markov kernel \mathbb{M} be given by equations (3.1) and (3.2) for the case that $(S, \mathfrak{S}) \subseteq (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Assume that \mathbb{M} arises from a full conditional distribution that is derived from a strictly positive joint density f w.r.t. the Lebesgue measure $\lambda^{|V|}$. Then the Markov kernel is ergodic.*

Proof. It suffices to verify that the requirements of the Aperiodic-Ergodic-Theorem are fulfilled, cf. Meyn and Tweedie [2009] Theorem 13.0.1. Plainly, the Markov kernel is $\lambda^{|V|d}$ -irreducible and $\lambda^{|V|d}$ is equivalent to any maximal irreducibility measure. Furthermore, since f is strictly positive, for any $B \in \mathfrak{S}^{\otimes |V|}$ with positive Lebesgue measure, $\mathbb{M}(x, B) > 0$ for all $x \in S^{|V|}$. Hence, \mathbb{M} is aperiodic. By Theorem 3.4 the existence of an invariant probability measure is fulfilled. By Theorem 10.1.1 and 10.0.1 in Meyn and Tweedie [2009] this invariant probability measure is unique. Furthermore, for each $x \in S$ the probability measure $\mathbb{M}(x, \cdot)$ is clearly absolutely continuous with respect to the Lebesgue measure $\lambda^{|V|d}$ which again is equivalent to the stationary measure $\mathbb{P}_Y = \int_{\bullet} f d\lambda^{|V|d}$ on \mathfrak{S} . Thus, the requirements of Theorem 1.3 from Hernández-Lerma and Lasserre [2001] are met and the Markov chain is positive Harris recurrent and we can conclude from the Aperiodic-Ergodic-Theorem that \mathbb{M} is ergodic. \square

We give an example

Example 3.6 (Concliques and the normal distribution). Let $G = (V, E)$ be a finite graph and $\{Y(v) : v \in V\}$ be multivariate normal with expectation $\alpha \in \mathbb{R}^{|V|}$ and covariance $\Sigma \in \mathbb{R}^{|V| \times |V|}$ in that Y has the density

$$f_Y(y) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y - \alpha)^T \Sigma^{-1} (y - \alpha) \right\}.$$

Then for a node v we have using the notation P for the precision matrix Σ^{-1}

$$Y(v) | Y(-v) \sim \mathcal{N} \left(\alpha(v) - (P(v, v))^{-1} \sum_{w \neq v} P(v, w) (Y(w) - \alpha(w)), (P(v, v))^{-1} \right).$$

Since $P = \Sigma^{-1}$ is symmetric and since we can assume that $(P(v, v))^{-1} > 0$, Y is a Markov random field if and only if for all nodes $v \in V$

$$P(v, w) \neq 0 \text{ for all } w \in Ne(v) \text{ and } P(v, w) = 0 \text{ for all } w \in V \setminus Ne(v).$$

Cressie [1993] investigates the conditional specification

$$Y(v) | Y(-v) \sim \mathcal{N} \left(\alpha(v) + \sum_{w \in Ne(v)} c(v, w) (Y(w) - \alpha(w)), \tau^2(v) \right)$$

where $C = (c(v, w))_{v, w}$ is a $|V| \times |V|$ matrix and $T = \text{diag}(\tau^2(v) : v \in V)$ is a diagonal matrix such that the coefficients satisfy the necessary condition $\tau^2(v)c(w, v) = \tau^2(w)c(v, w)$ for $v \neq w$ and $c(v, v) = 0$ as well as $c(v, w) = 0 = c(w, v)$ if v, w are no neighbors. This means $P(v, w) = -c(v, w)P(v, v)$, i.e., $\Sigma^{-1} = P = T^{-1}(I - C)$. If $I - C$ is invertible and $(I - C)^{-1}T$ is symmetric and positive definite, then the entire random field is multivariate normal with $Y \sim \mathcal{N}(\alpha, (I - C)^{-1}T)$.

With this insight it is possible to simulate a Gaussian Markov random field using concliques with a consistent full conditional distribution. In particular, it is plausible in many applications to use equal weights $c(v, w)$ (cf. Cressie [1993]): we can write the matrix C as $C = \eta H$ where H is the adjacency matrix of G , i.e., $H(v, w)$ is 1 if v, w are neighbors, otherwise it is 0. We know from the properties of the Neumann series that $I - C$ is invertible if $(h_0)^{-1} < \eta < (h_m)^{-1}$ where h_m is the maximal and h_0 is the minimal eigenvalue of H ,

3.2. Numerical results. In Example 3.6 we have considered the multivariate normal distribution in the context of Markov random fields on a finite graph. We continue with this idea at this point: let $G = (V, E)$ be a finite graph with nodes $v_1, \dots, v_{|V|}$, we simulate a d -dimensional random field Z on G such that each component Z_i takes values in $\mathbb{R}^{|V|}$, $i = 1, \dots, d$. Here we use copulas to simulate some of the components Z_i as dependent. Each random field Z_i has a specification

$$Z_i \sim \mathcal{N}(\alpha(1, \dots, 1)', \sigma^2 \Sigma) \quad (3.3)$$

where $\alpha, \sigma \in \mathbb{R}$ and $\sigma > 0$; furthermore, Σ is a correlation matrix which satisfies the relation

$$(I - \eta H)^{-1} T = \sigma^2 \Sigma. \quad (3.4)$$

The matrix H is the adjacency matrix of G . The parameter η is chosen such that $I - \eta H$ is invertible and T is a diagonal matrix $T = \text{diag}(\tau^2(v_1), \dots, \tau^2(v_{|V|}))$. A large absolute value of η indicates a strong dependence within the random variables of one component, whereas $\eta = 0$ indicates independence within the component. The marginal laws within a component are equal: $Z_i(v) \sim \mathcal{N}(\alpha, \sigma^2)$ for $v \in V$. However, the conditional variances $\tau^2(\cdot)$ within a component Z_i may differ.

In the next step, we construct from some components the random field $\{X(v) : v \in V\}$ and from another

independent component the error terms $\{\varepsilon(v) : v \in V\}$, we precise this below. For a choice m as conditional mean function and a constant conditional variance function ς , we then simulate the field Y as in equation (1.1) and estimate m with the least-squares estimator from equations (1.2) and (1.4). In the situation where the regression function m is known, the L^2 -error can serve as a criterion for the goodness-of-fit of the estimate for m given \hat{m} : we split the whole sample into a learning sample V_L and a testing sample V_T . Here both V_L and V_T should be two connected sets w.r.t. the underlying graph if this is possible. We estimate \hat{m} from the learning sample and compute the approximate L^2 -error with Monte Carlo integration over the testing sample, i.e.,

$$\int_{\mathbb{R}^d} |\hat{m} - m|^2 d\mu_X \approx |V_T|^{-1} \sum_{v \in V_T} |\hat{m}(X(v)) - m(X(v))|^2.$$

In order to obtain the distributional characteristics of the L^2 -error, we repeat this whole procedure $M_1 = 1000$ times.

Example 3.7 (Bivariate nonparametric regression on Gaussian Markov random fields). We simulate a random field on a planar graph $G = (V, E)$ that represents the administrative divisions in the Sidney bay area on the statistical area level 1 (for further reference, compare the webpage of the Australian bureau of statistics, www.abs.gov.au). It comprises 7,713 nodes and approximately 47k edges in total. Hence, G is highly connected if compared to the standard four-nearest neighborhood lattice. An illustration of the graph is given in Figure 1a. On this graph we model a 3-dimensional Gaussian Markov random field $Z = (Z_1, Z_2, Z_3)$ each having a specification as in Example 3.6 such that the marginals $Z_i(v)$ within each component are standard normally distributed. The parameter space for η is derived from the adjacency matrix of the graph G which we denote by H and which contains the interval $(-0.2221, 0.1312)$. Mark that the range for the lattice with a four-nearest-neighborhood structure is $(-0.25, 0.25)$. The marginal conditional variance of the variable $Z_i(v)$ which is given by $\tau_i^2(v)$ is then adjusted such that the entire random vector Z_i has a covariance structure of the type Σ_i as in (3.4) for a correlation matrix Σ_i for $i = 1, 2, 3$.

	Estimates on the graph		Independent reference estimates	
j	D4 wavelet	Haar wavelet	D4 wavelet	Haar wavelet
1	0.264 (0.006)	0.413 (0.008)	0.260 (0.006)	0.406 (0.007)
2	0.122 (0.009)	0.258 (0.008)	0.119 (0.009)	0.254 (0.007)
3	0.163 (0.036)	0.198 (0.010)	0.170 (0.044)	0.196 (0.010)
4	0.422 (0.075)	0.259 (0.012)	0.435 (0.077)	0.257 (0.012)

TABLE 1. L^2 -error of the bivariate regression problem: the estimated mean and in brackets the estimated standard deviation for a resolution $j = 1, \dots, 4$. The first two columns give the results for the random field, the last two columns those of the independent reference sample.

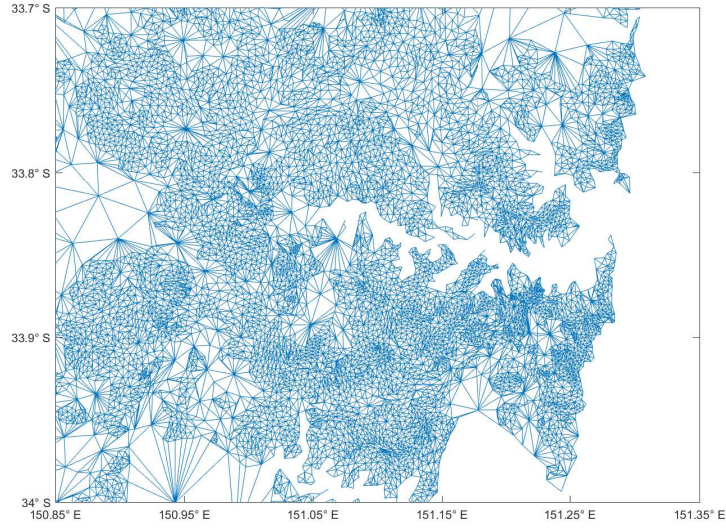
In order to obtain dependent components Z_1 and Z_2 , we simulate these with Algorithm 3.3 and draw the error terms from a 2-dimensional Gaussian copula in each iteration. The exact simulation parameters are given by

$$\mu_{Z_i} = 0, \sigma_i = 1 \text{ for } i = 1, 2, 3, \eta_1 = 0.12, \eta_2 = -0.18 \text{ and } \eta_3 = 0.12.$$

The covariance between the first two components is 0.7. The third component Z_3 is simulated as independent. The vectors $\tau_i^2 \in \mathbb{R}^{|V|}$ ($i = 1, 2, 3$) are computed with the formula $\tau_i^2(v) = \left\{ \text{diag}(\text{inv}(I - \eta_i H)) \right\}^{-1}(v)$, where we denote here by inv the inverse of a matrix, by diag the operator that maps the diagonal of a matrix to a vector and by $\{\cdot\}^{-1}$ the elementwise inversion of a vector. Afterwards, we transform the first two components Z_1 and Z_2 with a two dimensional standard normal distribution onto the unit square and obtain the random field (X_1, X_2) . For the random field Y we specify the following mean function

$$m : \mathbb{R}^2 \rightarrow \mathbb{R}, (x_1, x_2) \mapsto (2 - 3x_2^2 + 4x_2^4) \exp(-(2x_1 - 1)^2).$$

The function plot of m is given in Figure 1b. We simulate $Y(v) = m(X_1(v), X_2(v)) + Z_3(v)$, hence, the conditional variance ς is constant and equal to 1. We run $M_2 = 15k$ iteration steps in the Markov chain algorithm 3.3. We



(A) The Sidney bay area (on statistical area level 1-scale)

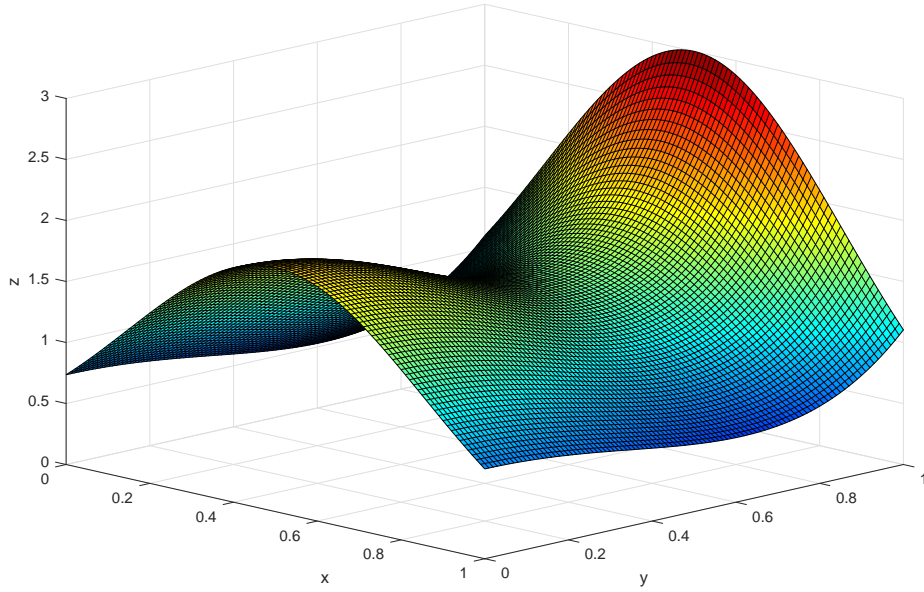
(B) Function plot of m

FIGURE 1. Input graph and regression function for the bivariate regression problem

use two different wavelet scaling functions for the estimation of m with nonparametric regression: we run the first regression with the Haar wavelet scaling function $\varphi = 1_{[0,1]}$ and the second regression with Daubechies 4-scaling function $D4$ (resp. $db2$). We proceed in each case as follows: given the domain of X , which in this case is $[0, 1]^2$, we fix a resolution scale j which rescales the domain of the scaling function φ by the factor 2^{-j} , then we select all integer combinations $\gamma_i = (\gamma_{i,1}, \gamma_{i,2})^t$ for which the domain of the 2-dimensional scaling function $\Phi = \varphi \otimes \varphi$ when rescaled and shifted by γ intersects with $[0, 1]^2$. We get a list of n functions

$$f_i = \varphi(2^{-j} \cdot -\gamma_{i,1}) \otimes \varphi(2^{-j} \cdot -\gamma_{i,2}), \quad i = 1, \dots, n$$

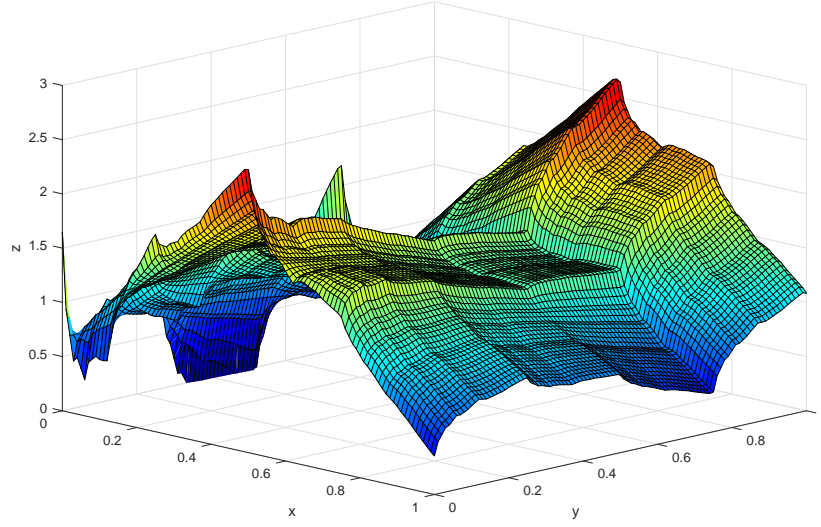
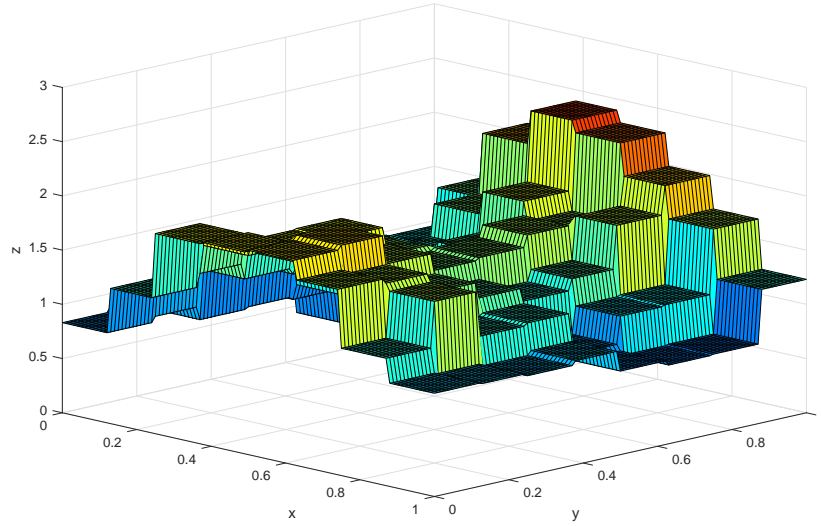
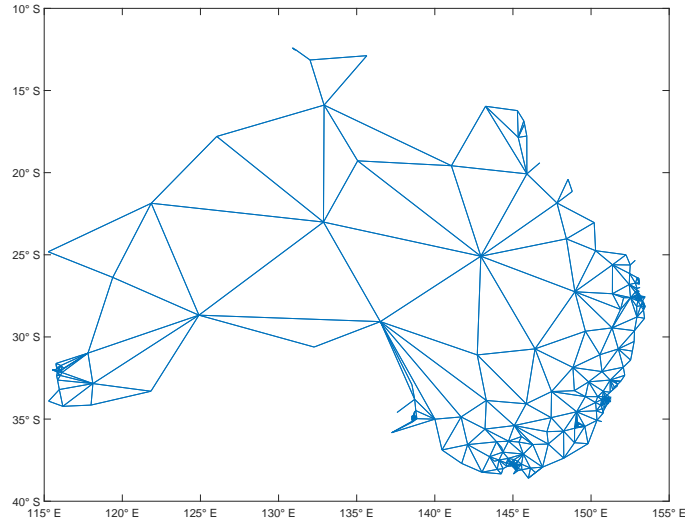
(A) Estimate \hat{m} with the D4 scaling function and the scaling parameter $j = 2$ (B) Estimate \hat{m} with the Haar scaling function and the scaling parameter $j = 3$

FIGURE 2. Estimated regression functions for a bivariate regression problem

which we evaluate at the data \tilde{X} and obtain a regression matrix. We use a singular value decomposition of this regression matrix to perform the standard least-squares procedure which yields the coefficients $c_i, i = 1, \dots, n$, for the functions f_i such that $\hat{m} = \sum_{i=1}^n c_i f_i$. The results are given in Figure 2: Figure 2a depicts the estimates based on Daubechies 4-scaling function, Figure 2b those based on the Haar scaling function, Daubechies 4-scaling function outperforms slightly the Haar-scaling function in this case.

Example 3.8 (Univariate nonparametric regression on Gaussian Markov random fields). In this example we consider a one dimensional spatial regression problem based on a graph which represents Australia when divided into administrative division on the statistical area level 3. The graph consists of 330 nodes and 1600 edges, cf. Figure 3a; hence, again this graph is highly connected.

We simulate two Gaussian random fields Z_1 and Z_2 on G with marginal means 0 and marginal variances 1 with the Markov chain method as in Example 3.7. The parameter space for η contains the interval $(-0.3060, 0.1615)$,



(A) Administrative divisions of mainland Australia

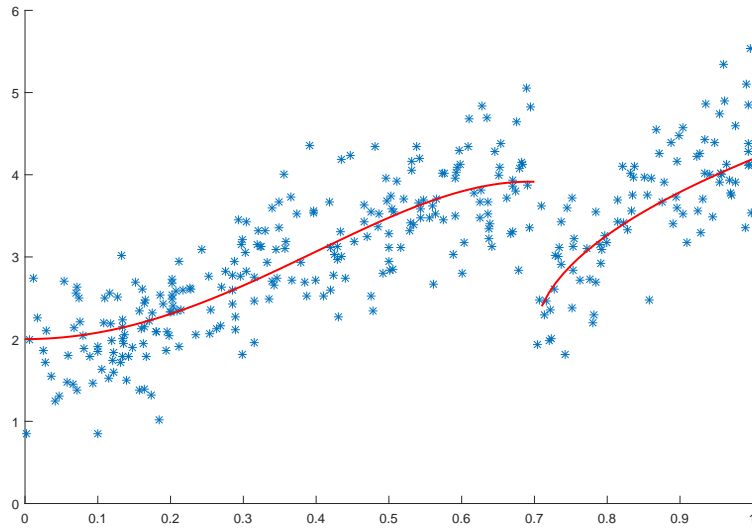
(B) A realization of X and the mean function m

FIGURE 3. Graph and true regression function.

we choose η for both components equal to 0.15. We run $M_2 = 15k$ simulations. Then we retransform the component Z_1 on the unit interval with an inverse standard normal distribution and obtain the random field X whose marginals are approximately uniformly distributed on $[0, 1]$. The conditional mean function is given by the noncontinuous function

$$m : [0, 1] \rightarrow \mathbb{R}, x \mapsto (2 + 8x^2 - (1.7x)^4) 1_{\{x \leq 0.7\}} + 2(\sqrt{4(x - 0.7)} + 1) 1_{\{0.7 < x\}}.$$

We specify Y as $Y(v) = m(X(v)) + Z_2(v)/2$. Figure 3b depicts the simulated random field. Figure 4a shows the estimation with the Daubechies 4-scaling function, while 4b depicts the case for Haar wavelet. Table 2 shows that the L^2 -error is minimized in all cases for the resolution $j = 4$. Mark that in this example Daubechies wavelet consistently outperforms the Haar wavelet when measured by the theoretic L^2 -error.

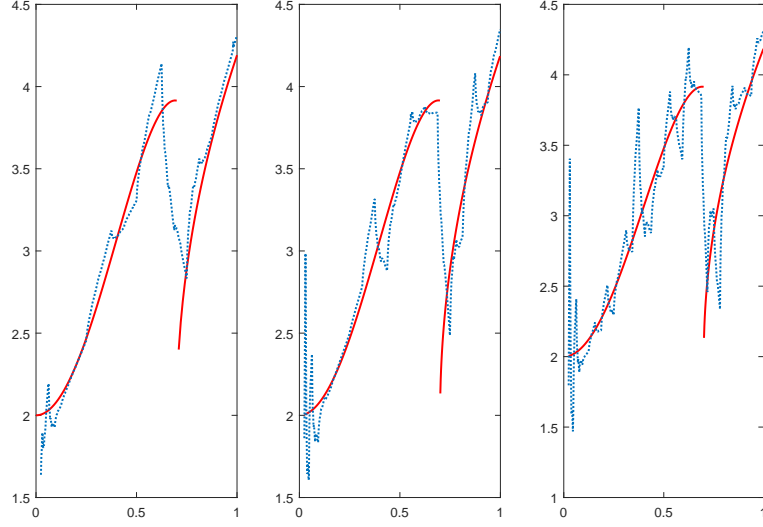
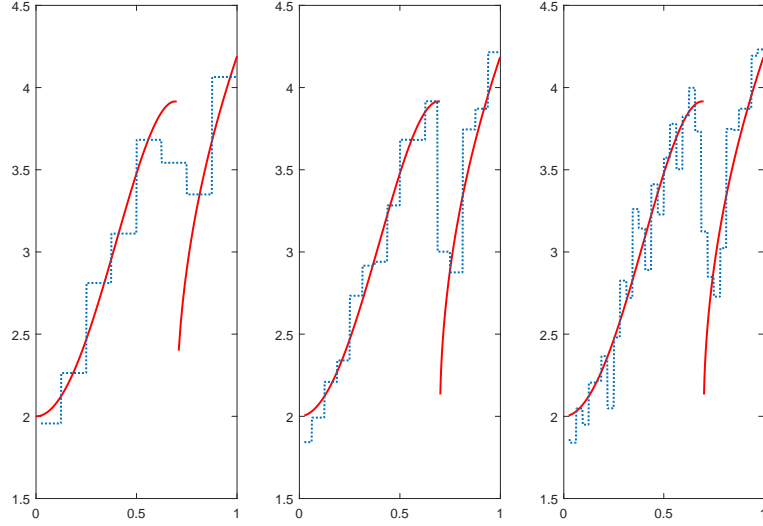
(A) D4 estimate for $j = 3, 4, 5$.(B) Haar estimate for $j = 3, 4, 5$.

FIGURE 4. The estimates for the univariate regression problem.

4. PROOFS OF THE THEOREMS IN SECTION 1 AND SECTION 2

The upcoming proposition is a well-known result due to Györfi et al. [2002] which states sufficient conditions for a consistent estimator.

Proposition 4.1 (Modified version of Györfi et al. [2002] Theorem 10.2). *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space endowed with the random field $(X, Y) := \{(X(s), Y(s)) : s \in I\}$ from equation (1.1) where each $X(s)$ is \mathbb{R}^d -valued and each $Y(s)$ is \mathbb{R} -valued. Let (X, Y) satisfy Condition 1.4. Let $Y(s)$ be square integrable and denote by μ_X the marginal law of the $X(s)$. For each $k \in \mathbb{N}_+$ let $\mathcal{F}_k \subseteq L^2(\mu_X)$ be a deterministic class of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$. Denote by $T_{\beta_k} \mathcal{F}_k$ the truncated function classes and by \hat{m}_k the truncated least-squares estimate of m given in equations (1.2) and (1.4) for some sequence $\{\beta_k : k \in \mathbb{N}\}$ increasing to infinity. In addition, let*

		Estimates on the graph		Independent reference estimates	
j		D4 wavelet	Haar wavelet	D4 wavelet	Haar wavelet
2		0.326 (0.031)	0.405 (0.059)	0.321 (0.029)	0.401 (0.061)
3		0.241 (0.033)	0.344 (0.064)	0.233 (0.035)	0.341 (0.067)
4		0.224 (0.077)	0.284 (0.073)	0.213 (0.062)	0.280 (0.078)
5		0.319 (0.172)	0.349 (0.117)	0.299 (0.134)	0.333 (0.093)
6		0.772 (0.437)	0.753 (0.213)	0.712 (0.380)	0.727 (0.212)

TABLE 2. L^2 -error of the univariate regression problem: the estimated mean and in brackets the estimated standard deviation for a resolution $j = 2, \dots, 6$. The first two columns give the results for the random field, the last two columns those of the independent reference sample.

the positive valued mapping

$$\Omega \ni \omega \mapsto \sup_{f \in T_{\beta_k} \mathcal{F}_k} \left| \frac{1}{|I_{n(k)}|} \sum_{s \in I_{n(k)}} (T_L Y(s, \omega) - f(X(s, \omega)))^2 - \mathbb{E} \left[(T_L Y(e_N) - f(X(e_N)))^2 \right] \right|$$

be \mathcal{A} -measurable.

(a) If for all $L > 0$ both

$$\lim_{k \rightarrow \infty} \mathbb{E} \left[\inf_{\substack{f \in \mathcal{F}_k, \\ \|f\|_{\infty} \leq \beta_k}} \|f - m\|_{L^2(\mu_X)} \right] = 0 \text{ and}$$

$$\lim_{k \rightarrow \infty} \mathbb{E} \left[\sup_{f \in T_{\beta_k} \mathcal{F}_k} \left| \frac{1}{|I_{n(k)}|} \sum_{s \in I_{n(k)}} (T_L Y(s) - f(X(s)))^2 - \mathbb{E} \left[(T_L Y(e_N) - f(X(e_N)))^2 \right] \right| \right] = 0,$$

then, $\{\hat{m}_k : k \in \mathbb{N}_+\}$ is weakly consistent in that

$$\lim_{k \rightarrow \infty} \mathbb{E} \left[\int_{\mathbb{R}^d} (\hat{m}_k(z) - m(z))^2 \mu_X(dz) \right] = 0.$$

(b) If, furthermore, $\frac{1}{|I_{n(k)}|} \sum_{s \in I_{n(k)}} |Y(s) - T_L Y(s)|^2 \rightarrow \mathbb{E} [|Y(e_N) - T_L Y(e_N)|^2]$ a.s. and if both

$$\lim_{k \rightarrow \infty} \inf_{\substack{f \in \mathcal{F}_k, \\ \|f\|_{\infty} \leq \beta_k}} \|f - m\|_{L^2(\mu_X)} = 0 \quad \text{a.s. and}$$

$$\lim_{k \rightarrow \infty} \sup_{f \in T_{\beta_k} \mathcal{F}_k} \left| \frac{1}{|I_{n(k)}|} \sum_{s \in I_{n(k)}} (T_L Y(s) - f(X(s)))^2 - \mathbb{E} \left[(T_L Y(e_N) - f(X(e_N)))^2 \right] \right| = 0 \quad \text{a.s.}$$

for all $L > 0$, then $\{\hat{m}_k : k \in \mathbb{N}_+\}$ is strongly consistent in that

$$\lim_{k \rightarrow \infty} \int_{\mathbb{R}^d} (\hat{m}_k(z) - m(z))^2 \mu_X(dz) = 0 \quad \text{a.s.}$$

It follows the proof of the first main theorem of Section 1

Proof of Theorem 1.7. We verify that in both cases the sufficient criteria in Proposition 4.1 are satisfied. The structure of the proof is identical to that of Theorem 10.3 in Györfi et al. [2002], what differs are the bounds. Therefore we sketch the major parts. W.l.o.g. we can assume that $L < \beta_k$. We have to consider the function classes (for $k \in \mathbb{N}_+$)

$$\mathcal{H}_k := \left\{ h : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}, h(x, y) = |f(x) - T_L(y)|^2 \right.$$

for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, for some $f \in T_{\beta_k} \mathcal{F}_k$.

Denote by $H_{\mathcal{H}_k}(\varepsilon)$ a uniform bound on the ε -covering number $\mathcal{N}(\varepsilon, \mathcal{H}_k, \|\cdot\|_{L^1(\nu)})$ where ν is an arbitrary probability measure with equal masses on the points $z_1, \dots, z_u \in \mathbb{R}$, $u \in \mathbb{N}_+$. For this very class \mathcal{H}_k we have, provided that $L \leq \beta_k$, and under Condition 1.6

$$H_{\mathcal{H}_k}\left(\frac{\varepsilon}{32}\right) \leq H_{T_{\beta_k} \mathcal{F}_k}\left(\frac{\varepsilon}{32(4\beta_k)}\right) = H_{T_{\beta_k} \mathcal{F}_k}\left(\frac{\varepsilon}{128\beta_k}\right) = \exp \kappa_k(\varepsilon, \beta_k).$$

Note that the functions in \mathcal{H}_k are bounded by $4\beta_k^2$ if $L \leq \beta_k$. By assumption

$$\beta_k^2 \kappa_k(\varepsilon, \beta_k) \left(\prod_{i=1}^N \log n_i(k) \right) / \left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)} \rightarrow 0 \text{ as } k \rightarrow \infty,$$

thus, Theorem A.5 reduces to

$$\begin{aligned} & \mathbb{P} \left(\sup_{f \in T_{\beta_k} \mathcal{F}_k} \left| \frac{1}{|I_{n(k)}|} \sum_{s \in I_{n(k)}} |f(X(s)) - T_L Y(s)|^2 - \mathbb{E} \left[|f(X(e_N)) - T_L Y(e_N)|^2 \right] \right| > \varepsilon \right) \\ & \leq A_1 \exp \{ \kappa_k(\varepsilon, \beta_k) \} \exp \left\{ - \frac{A_2 \varepsilon \left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)}}{\beta_k^2 \prod_{i=1}^N \log n_i(k)} \right\} \\ & = A_1 \exp \left\{ - \frac{\left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)}}{\beta_k^2 \prod_{i=1}^N \log n_i(k)} \left(A_2 \varepsilon - \frac{\beta_k^2 \kappa_k(\varepsilon, \beta_k) \prod_{i=1}^N \log n_i(k)}{\left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)}} \right) \right\} \end{aligned} \quad (4.1)$$

for suitable constants A_1 and A_2 . The weak consistency follows from (4.1). Indeed,

$$\begin{aligned} & \mathbb{E} \left[\sup_{f \in T_{\beta_n} \mathcal{F}_n} \left| \frac{1}{|I_n|} \sum_{s \in I_n} |f(X(s)) - T_L Y(s)|^2 - \mathbb{E} \left[|f(X(e_N)) - T_L Y(e_N)|^2 \right] \right| \right] \\ & \leq \varepsilon + A_1 \exp \{ \kappa_k(\varepsilon, \beta_k) \} \int_{\varepsilon}^{\infty} \exp \left\{ - \frac{A_2 t \left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)}}{\beta_k^2 \prod_{i=1}^N \log n_i(k)} \right\} dt \\ & \leq \varepsilon + \frac{A_1}{A_2} \frac{\beta_k^2 \prod_{i=1}^N \log n_i(k)}{\left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)}} \\ & \quad \cdot \exp \left\{ - \frac{\left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)}}{\beta_k^2 \prod_{i=1}^N \log n_i(k)} \left(A_2 \varepsilon - \frac{\beta_k^2 \kappa_k(\varepsilon, \beta_k) \prod_{i=1}^N \log n_i(k)}{\left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)}} \right) \right\} \rightarrow \varepsilon, \end{aligned}$$

as $k \rightarrow \infty$. Furthermore, if additionally for some $\delta > 0$,

$$\left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)} / \left\{ \beta_k^2 \left(\prod_{i=1}^N \log n_i(k) \right) (\log k)^{1+\delta} \right\} \rightarrow \infty \text{ as } k \rightarrow \infty,$$

equation (4.1) remains summable over k . Now an application of the Borel-Cantelli Lemma to the same equation and the requirement that

$$\lim_{k \rightarrow \infty} \frac{1}{|I_{n(k)}|} \sum_{s \in I_{n(k)}} |Y(s) - T_L Y(s)|^2 = \mathbb{E} \left[|Y(e_N) - T_L Y(e_N)|^2 \right] \text{ a.s.}$$

for all $L > 0$ yield that the estimator is strongly universally consistent. This finishes the proof. \square

Proof of Corollary 1.8. Clearly, the map

$$\mathbb{R}^{K_k} \times \Omega \ni (a, \omega) \mapsto \sum_{i=1}^{K_k} a_i f_i(X(s, \omega)) \text{ is } \mathcal{B}(\mathbb{R}^{K_k}) \otimes \mathcal{A}\text{-measurable.}$$

The desired measurability from equation (1.5) follows now from the fact that for any measurable function g on a product space $(S \times T, \mathcal{S} \otimes \mathcal{T})$ the set

$$\begin{aligned} \left\{ t \in T : \sup_{s \in S} g(s, t) > c \right\} &= \left\{ t \in T \mid \exists s \in S : g(s, t) > c \right\} \\ &= \pi_T^{S \times T} \{ (s, t) \in S \times T : g(s, t) > c \} \in \mathcal{T}, \end{aligned}$$

where we denote by $\pi_T^{S \times T}$ the projection from $S \times T$ onto T .

Furthermore, the Vapnik-Chervonenkis-dimension is at least 2 if $K_k \geq 2$. Indeed, choose f_1 and f_2 . Without loss of generality, there is an \bar{x} in \mathbb{R}^d and an a in \mathbb{R} such that $af_1(\bar{x}) = f_2(\bar{x}) > 0$. Since f_1 and f_2 are linear independent exactly one of the three cases occurs: (1) either there are x_1 and x_2 in a neighborhood of \bar{x} such that $af_1(x_1) > f_2(x_1)$ and $f_2(x_2) > af_1(x_2)$, (2) or $af_1 = f_2$ on U and $af_1 > f_2$ on $\mathbb{R}^d \setminus U$, where $U \subset \mathbb{R}^d$ contains \bar{x} , (3) or $f_2 = af_1$ on U and $f_2 > af_1$ on $\mathbb{R}^d \setminus U$. In the last two cases we can modify a by some amount such that we achieve the first case, by linear independence. Thus, the two points $p_i := (x_i, t_i)$ ($i=1,2$) with the property that $af_1(x_1) > t_1 > f_2(x_1)$ and $f_2(x_2) > t_2 > af_1(x_2)$ are shattered by the set of all subgraphs of the linear space $\langle f_1, f_2 \rangle$, hence, $\mathcal{V}_{\langle f_1, \dots, f_n \rangle^+} \geq \mathcal{V}_{\langle f_1, f_2 \rangle^+} \geq 2$. Consequently, the conditions of Theorem A.1 are fulfilled. We have

$$\begin{aligned} \kappa_k(\varepsilon, \beta_k) &= \log H_{T_{\beta_k} \mathcal{F}_k} \left(\frac{\varepsilon}{128\beta_k} \right) \leq \log \left(3 \left(\frac{512e\beta_k^2}{\varepsilon} \log \frac{768e\beta_k^2}{\varepsilon} \right)^{\mathcal{V}_{(T_{\beta_k} \mathcal{F}_k)^+}} \right) \\ &\leq (K_k + 1) \log \left(3(768)^2 \left(\frac{e}{\varepsilon} \right)^2 \beta_k^4 \right). \end{aligned}$$

In addition, in this case the variables $\{|Y(s) - T_L Y(s)|^2 : s \in \mathbb{Z}^N\}$ are ergodic, cf. Theorem B.3, which implies that $\frac{1}{|I_n(k)|} \sum_{s \in I_n(k)} |Y(s) - T_L Y(s)|^2 \rightarrow \mathbb{E} [|Y(e_N) - T_L Y(e_N)|^2]$ a.s. for all $L > 0$. This finishes the proof. \square

We introduce for notational convenience

Notation 4.2. Let f be a real valued function on \mathbb{R}^d and let the stationary distribution of the $X(s)$ be given by μ_X . We write $\|f\| := \left(\int_{\mathbb{R}^d} f^2 d\mu_X \right)^{\frac{1}{2}}$ for the $L^2(\mu_X)$ -norm. Furthermore, let a sample $\{X(s) : s \in J\}$ from a random field be given where $J \subseteq \mathbb{N}_+^N$ is finite as well as an i.i.d. ghost sample $\{X'(s) : s \in J\}$ with the same marginals as X . Define the following empirical L^2 -norms (w.r.t. J)

$$\begin{aligned} \|f\|_J &:= \left(\frac{1}{|J|} \sum_{s \in J} f(X(s))^2 \right)^{\frac{1}{2}}, \quad \|f\|'_J := \left(\frac{1}{|J|} \sum_{s \in J} f(X'(s))^2 \right)^{\frac{1}{2}} \\ \text{and } \|f\|_{\tilde{J}} &:= \left(\frac{1}{2|J|} \sum_{s \in J} f(X(s))^2 + f(X'(s))^2 \right)^{\frac{1}{2}}. \end{aligned}$$

The following proposition prepares the second main theorem of Section 1, Theorem 1.9

Proposition 4.3. Let $\{X(s) : s \in I^+\}$ be random field that satisfies Condition 1.4. Let \mathcal{G} be a class of \mathbb{R} -valued functions on \mathbb{R}^d all bounded by a universal constant $0 < B < \infty$. Then, if the index set I_n fulfills both $\min_{1 \leq i \leq N} n_i \geq e^2$ and $|I_n| \geq 64B^2/\varepsilon^2$ is sufficiently large,

$$\begin{aligned} &\mathbb{P} \left(\sup_{f \in \mathcal{G}} \|f\| - 2\|f\|_{I_n} > \varepsilon \right) \\ &\leq A_1 \left\| \mathcal{N}_2 \left(\frac{\varepsilon}{16\sqrt{2}}, \mathcal{G}, (X(I_n), X'(I_n)) \right) \right\|_{\infty} \\ &\quad \cdot \left\{ \exp \left(-A_2 \varepsilon^2 \frac{\left(\prod_{i=1}^N n_i \right)^{\rho - N/(N+1)}}{B^2 \prod_{i=1}^N \log n_i} \right) + \exp \left(-A_3 \varepsilon^4 \frac{\left(\prod_{i=1}^N n_i \right)^{\rho}}{B^4} \right) \right\}, \end{aligned}$$

for constants $0 < A_1, A_2, A_3 < \infty$ which do not depend on the bound B nor on ε nor on the index set I_n .

Note that under the assumption that $\mathcal{V}_{\mathcal{G}^+} \geq 2$ and ε sufficiently small the bound from Proposition 4.3 is non-trivial, by Theorem A.1 we have

$$\left\| \mathcal{N}_2 \left(\frac{\varepsilon}{16\sqrt{2}}, \mathcal{G}, (X(I_n), X'(I_n)) \right) \right\|_{\infty} \leq 3 \left(\frac{16^3 e B^2}{\varepsilon^2} \cdot \log \frac{24 \cdot 16^2 e B^2}{\varepsilon^2} \right)^{\mathcal{V}_{\mathcal{G}^+}}.$$

Proof of Proposition 4.3. Let $\{X(s) : s \in I_n\}$ be a subset of the strong mixing and stationary random field X and let $\{X'(s) : s \in I_n\}$ be the corresponding ghost sample. One can show that

$$\mathbb{P}\left(\exists f \in \mathcal{G} : \|f\| - 2\|f\|_{I_n} > \varepsilon\right) \leq \frac{3}{2} \mathbb{P}\left(\exists f \in \mathcal{G} : \|f\|'_{I_n} - \|f\|_{I_n} > \frac{\varepsilon}{4}\right)$$

if $|I_n| \geq 64B^2/\varepsilon^2$, cf. Györfi et al. [2002] proof of Theorem 11.2. This relation holds in the same way for a dependent array of random variables with equal marginal distributions. In the next step, we consider things for each $\omega \in \Omega$ separately. Let U_1, \dots, U_{H^*} be a $\varepsilon/(16\sqrt{2})$ -covering of \mathcal{G} with respect to the empirical L^2 -norm of the whole sample $(X(I_n), X'(I_n))$ with the notation $H^* := \mathcal{N}_2(\varepsilon/(16\sqrt{2}), \mathcal{G}, (X(I_n), X'(I_n)))$ and $U_k := \{f \in \mathcal{G} : \|f - g_k\|'_{I_n} < \varepsilon/(16\sqrt{2})\}$, where the covering functions are g_1, \dots, g_{H^*} . Note that H^* and the U_k are random and that both $\|\cdot\|_{I_n}$ and $\|\cdot\|'_{I_n}$ are bounded by $\sqrt{2}\|\cdot\|'_{I_n}$. Then,

$$\mathbb{P}\left(\exists f \in \mathcal{G} : \|f\|'_{I_n} - \|f\|_{I_n} > \frac{\varepsilon}{4}\right) \leq \sum_{k=1}^{\|H^*\|_\infty} \mathbb{P}\left(\exists f \in U_k : \|f\|'_{I_n} - \|f\|_{I_n} > \frac{\varepsilon}{4}\right). \quad (4.2)$$

Now, we have for $f \in U_k$ and the fact that $\|f\|_{I_n} \leq \sqrt{2}\|f\|'_{I_n}$ the inequality

$$\begin{aligned} \|f\|'_{I_n} - \|f\|_{I_n} &= \|f\|'_{I_n} - \|g_k\|'_{I_n} + \|g_k\|'_{I_n} - \|g_k\|_{I_n} + \|g_k\|_{I_n} - \|f\|_{I_n} \\ &\leq \|f - g_k\|'_{I_n} + (\|g_k\|'_{I_n} - \|g_k\|_{I_n}) + \|f - g_k\|_{I_n} \\ &\leq 2\sqrt{2} \frac{\varepsilon}{16\sqrt{2}} + (\|g_k\|'_{I_n} - \|g_k\|_{I_n}). \end{aligned}$$

Hence, $\{\exists f \in U_k : \|f\|'_{I_n} - \|f\|_{I_n} > \frac{\varepsilon}{4}\} \subseteq \{\|g_k\|'_{I_n} - \|g_k\|_{I_n} > \frac{\varepsilon}{8}\}$ and since for $a, b, c \geq 0$ the inequality $a - b > c$ implies $a^2 - b^2 > c^2$, we get for the probability in equation (4.2) the following bounds

$$\begin{aligned} \mathbb{P}\left(\|g_k\|'_{I_n} - \|g_k\|_{I_n} > \frac{\varepsilon}{8}\right) &\leq \mathbb{P}\left(\left(\|g_k\|'_{I_n}\right)^2 - \left(\|g_k\|_{I_n}\right)^2 > \frac{\varepsilon^2}{64}\right) \\ &\leq \mathbb{P}\left(\frac{1}{|I_n|} \sum_{s \in I_n} \{g_k(X'(s))^2 - \mathbb{E}[g_k(X'(e_N))^2]\} \right. \\ &\quad \left. - \frac{1}{|I_n|} \sum_{s \in I_n} \{g_k(X(s))^2 - \mathbb{E}[g_k(X(e_N))^2]\} > \frac{\varepsilon^2}{64}\right) \\ &\leq \mathbb{P}\left(\left|\frac{1}{|I_n|} \sum_{s \in I_n} g_k(X'(s))^2 - \mathbb{E}[g_k(X'(e_N))^2]\right| > \frac{\varepsilon^2}{128}\right) \\ &\quad + \mathbb{P}\left(\left|\frac{1}{|I_n|} \sum_{s \in I_n} g_k(X(s))^2 - \mathbb{E}[g_k(X(e_N))^2]\right| > \frac{\varepsilon^2}{128}\right) \end{aligned} \quad (4.3)$$

The first term from (4.3) can be bounded by Hoeffding's inequality, we have

$$\mathbb{P}\left(\left|\frac{1}{|I_n|} \sum_{s \in I_n} g_k(X'(s))^2 - \mathbb{E}[g_k(X'(e_N))^2]\right| > \frac{\varepsilon^2}{128}\right) \leq 2 \exp\left(-\frac{C\varepsilon^4}{2^{15}} \frac{(\prod_{i=1}^N n_i)^p}{B^4}\right).$$

For the second term we get with Proposition A.4 that

$$\mathbb{P}\left(\left|\frac{1}{|I_n|} \sum_{s \in I_n} g_k(X(s))^2 - \mathbb{E}[g_k(X(s))^2]\right| > \frac{\varepsilon^2}{128}\right) \leq A_1 \exp\left(-A_2 \varepsilon^2 \frac{(\prod_{i=1}^N n_i)^{\rho-N/(N+1)}}{B^2 \prod_{i=1}^N \log n_i}\right),$$

for real constants A_1 and A_2 . This finishes the proof. \square

Proof of Theorem 1.9. We make use of the decomposition

$$\begin{aligned} &\int_{\mathbb{R}^d} |\hat{m}_k(x) - m(x)|^2 d\mu_X \\ &= \|\hat{m}_k - m\|^2 = (\|\hat{m}_k - m\| - 2\|\hat{m}_k - m\|_{I_{n(k)}} + 2\|\hat{m}_k - m\|_{I_{n(k)}})^2 \\ &\leq 2 \max(\|\hat{m}_k - m\| - 2\|\hat{m}_k - m\|_{I_{n(k)}}, 0)^2 + 8(\|\hat{m}_k - m\|_{I_{n(k)}})^2 \end{aligned} \quad (4.4)$$

The exponentially decreasing mixing rates ensure that the norm of the conditional covariance matrix remains bounded and that we can use Theorem 11.1 of Györfi et al. [2002] even in the case where the error terms $\varepsilon(s)$ are not independent: there is a constant C_1 such that $\|Cov(Y(I_{n(k)}) | X(I_{n(k)}))\|_2 \leq C_1$ for all $k \in \mathbb{N}$. Indeed, we have for matrices the norm inequality $\|\cdot\|_2 \leq \sqrt{\|\cdot\|_1 \|\cdot\|_\infty}$. Furthermore, as the covariance matrix is symmetric, the ∞ - and the 1-norm are equal. We consider a line (resp. a column) of the covariance matrix that contains the conditional covariances of the $Y(s)$. By assumption, the error terms satisfy $\mathbb{E}[|\varepsilon(s)|^{2+\delta}] < \infty$ for some $\delta > 0$. We use Davydov's inequality from Appendix A.2 and the bound on the mixing coefficients, $\alpha(k) \leq \lambda_0 \exp(-\lambda_1 k)$, certain $\lambda_0, \lambda_1 \in \mathbb{R}_+$. We get

$$\begin{aligned} & \sum_{t \in I_n} |Cov(Y(s), Y(t) | X(I_{n(k)}))| \\ & \leq \|\varsigma\|_\infty^2 \sum_{t \in I_{n(k)}} |Cov(\varepsilon(s), \varepsilon(t))| \leq 10 \|\varsigma\|_\infty^2 \mathbb{E}[|\varepsilon(s)|^{2+\delta}]^{2/(2+\delta)} \sum_{t \in I_{n(k)}} \alpha(\|s - t\|_\infty)^{\delta/(2+\delta)} \\ & \leq 10 \|\varsigma\|_\infty^2 \lambda_0 \mathbb{E}[|\varepsilon(s)|^{2+\delta}]^{2/(2+\delta)} \sum_{d=0}^{\max_{1 \leq i \leq N} n_i(k)} \exp(-\lambda_1 \delta/(2+\delta)d) ((2d+1)^N - (2d-1)^N) \\ & \leq C_1 < \infty, \end{aligned}$$

for all $s \in I_{n(k)}$ and all k and a suitable constant $C_1 \in \mathbb{R}$. Hence,

$$\|Cov(Y(I_{n(k)}) | X(I_{n(k)}))\|_2 \leq C_1.$$

Thus, by Theorem 11.1 of Györfi et al. [2002] which is (after a slight modification) applicable to dependent data as well,

$$\mathbb{E}[\|\hat{m}_k - m\|_{n(k)}^2] \leq C_1 \frac{K_k}{\left(\prod_{i=1}^N n_i(k)\right)^\rho} + \inf_{f \in \mathcal{F}_k} \int_{\mathbb{R}^d} (f(x) - m(x))^2 \mu_X(dx). \quad (4.5)$$

We apply Proposition 4.3 to the first term of (4.4). Therefore we denote by C' the constant from Condition 1.4 which fulfills $|I_{n(k)}| \geq C' \left(\prod_{i=1}^N n_i(k)\right)^\rho$. We have provided that $C' \left(\prod_{i=1}^N n_i(k)\right)^\rho \geq 128L^2/u$ is large enough

$$\begin{aligned} & \mathbb{P}\left(2 \left\{ \max(\|\hat{m}_k - m\| - 2\|\hat{m}_k - m\|_{I_{n(k)}}, 0) \right\}^2 > u\right) \\ & \leq \mathbb{P}\left(\exists f \in T_L \mathcal{F}_k : \|f - m\| - 2\|f - m\|_{I_{n(k)}} > \sqrt{\frac{u}{2}}\right). \end{aligned}$$

Furthermore, with Proposition A.1 and the estimates $\mathcal{V}_{T_L \mathcal{F}_k}^+ \leq \mathcal{V}_{\mathcal{F}_k}^+ \leq K_k + 1$,

$$\begin{aligned} & \left\| \mathcal{N}_2\left(\frac{\sqrt{v/2}}{\sqrt{2}16}, T_L \mathcal{F}_k, (X(I_{n(k)}), X'(I_{n(k)}))\right) \right\|_\infty \\ & \leq 3 \left(\frac{8 \cdot 32^2 e L^2}{v} \cdot \log \frac{12 \cdot 32^2 e L^2}{v} \right)^{K_k+1} \in O\left(\left(\frac{L^2}{v}\right)^{2(K_k+1)}\right), \end{aligned}$$

provided $\sqrt{v/2}/(\sqrt{2}16) < L/2$, i.e., $v < 16^2 L^2$. Hence, we get with Proposition 4.3 for $v < 16^2 L^2$ and $C' \left(\prod_{i=1}^N n_i(k)\right)^\rho \geq 128L^2/v$ the result

$$\begin{aligned} & \mathbb{E}\left[2 \left\{ \max(\|\hat{m}_n - m\| - 2\|\hat{m}_n - m\|_{I_n}, 0) \right\}^2\right] \\ & \leq v + \int_v^\infty \mathbb{P}\left(2 \left\{ \max(\|\hat{m}_n - m\| - 2\|\hat{m}_n - m\|_{I_n}, 0) \right\}^2 > u\right) du \\ & \leq v + A_1 \left(\frac{L^2}{v}\right)^{2(K_k+1)} \int_v^\infty \exp\left(-A_2 u \frac{\left(\prod_{i=1}^N n_i(k)\right)^{\rho-N/(N+1)}}{L^2 \prod_{i=1}^N \log n_i(k)}\right) \\ & \quad + \exp\left(-A_3 u^2 \frac{\left(\prod_{i=1}^N n_i(k)\right)^\rho}{L^4}\right) du. \end{aligned} \quad (4.6)$$

The second integral can be bounded with the inequalities:

$$\int_v^\infty \exp(-au^2) du = \sqrt{\frac{\pi}{a}} \Phi(-\sqrt{2av}) \leq \sqrt{\frac{\pi}{4a}} e^{-av^2}, \text{ for } a > 0.$$

Thus, under the assumption that $K_k \left(\prod_{i=1}^N \log n_i(k) \right)^3 / \left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)} \rightarrow 0$, one finds that (4.6) is in $O\left(K_k \left(\prod_{i=1}^N \log n_i(k) \right)^3 / \left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)}\right)$. This implies together with equation (4.5) that for some $A \in \mathbb{R}_+$

$$\begin{aligned} & \mathbb{E} \left[2 \left\{ \max \left(\|\hat{m}_k - m\| - 2 \|\hat{m}_k - m\|_{I_{n(k)}}, 0 \right) \right\}^2 \right] + 8 \mathbb{E} \left[(\|\hat{m}_k - m\|_{I_{n(k)}})^2 \right] \\ & \leq A \frac{K_k \left(\prod_{i=1}^N \log n_i(k) \right)^3}{\left(\prod_{i=1}^N n_i(k) \right)^{\rho-N/(N+1)}} + 8 \inf_{f \in \mathcal{F}_k} \int_{\mathbb{R}^d} |f(x) - m(x)|^2 \mu_X(dx) \text{ for all } k \in \mathbb{N}_+. \end{aligned}$$

□

We come to the proofs of the theorems in Section 2. Firstly, we show how to derive an isotropic MRA from a one-dimensional MRA

Proof of Example 2.3. In the first step, we show that the conditions for an MRA are fulfilled. The spaces $\cup_{j \in \mathbb{Z}} U_j$ are dense: by definition, we have

$$U_j = \otimes_{i=1}^d U'_j = \left\langle f_1 \otimes \dots \otimes f_d : f_i \in U'_j \forall i = 1, \dots, d \right\rangle.$$

Note that the set of pure tensors $\langle g_1 \otimes \dots \otimes g_d : g_i \in L^2(\lambda) \rangle$ is dense in $L^2(\lambda^d)$. Hence, it only remains to show that we can approximate any pure tensor $g_1 \otimes \dots \otimes g_d$ by a sequence $(F_j \in U_j : j \in \mathbb{N}_+)$. Let $\varepsilon > 0$ and a pure tensor $g_1 \otimes \dots \otimes g_d \in L^2(\lambda^d)$ be given. Choose a sequence of pure tensors $(f_{i,j} : j \in \mathbb{N}_+)$ converging to g_i in $L^2(\lambda)$ for $i = 1, \dots, d$. Denote by $L := \sup \{ \|f_{i,j}\|_{L^2(\lambda)}, \|g_i\|_{L^2(\lambda)} : j \in \mathbb{Z}, i = 1, \dots, d \} < \infty$. Then

$$\|g_1 \otimes \dots \otimes g_d - f_{1,j} \otimes \dots \otimes f_{d,j}\|_{L^2(\lambda^d)}^2 \leq d^2 L^{2(d-1)} \max_{1 \leq i \leq d} \|g_i - f_{i,j}\|_{L^2(\lambda)}^2 \rightarrow 0 \text{ as } j \rightarrow \infty.$$

Furthermore, $\cap_{j \in \mathbb{Z}} U_j = \{0\}$: Let $f = \sum_{i=1}^n a_i f_{i,1} \otimes \dots \otimes f_{i,d}$ be an element of each U_j . Then each $f_{i,k}$ is an element of each U'_j for all j and, hence, zero. The scaling property is immediate, too. Indeed,

$$\begin{aligned} f \in U_j & \Leftrightarrow f = \sum_{i=1}^n a_i f_{i,1} \otimes \dots \otimes f_{i,d} \text{ and } f_{i,k} \in U'_j, \quad k = 1, \dots, d \\ & \Leftrightarrow f = \sum_{i=1}^n a_i f_{i,1} \otimes \dots \otimes f_{i,d} \text{ and } f_{i,k}(2^{-j} \cdot) \in U'_0 \Leftrightarrow f(M^{-j} \cdot) \in U_0. \end{aligned}$$

The functions $\{\Phi(\cdot - \gamma) : \gamma \in \Gamma\}$ form an orthonormal basis of U_0 . We have for $\gamma, \gamma' \in \mathbb{Z}^d$

$$\begin{aligned} & \int_{\mathbb{R}^d} \Phi(x - \gamma) \Phi(x - \gamma') dx = \int_{\mathbb{R}^d} \otimes_{k=1}^d \varphi(x_k - \gamma_k) \cdot \otimes_{k=1}^d \varphi(x_k - \gamma'_k) dx \\ & = \prod_{k=1}^d \int_{\mathbb{R}} \varphi(x_k - \gamma_k) \varphi(x_k - \gamma'_k) dx_k = \delta_{\gamma, \gamma'} \end{aligned}$$

and for each $f \in U_0$ by definition $f = \sum_{i=1}^n a_i \varphi(\cdot - \gamma_1^i) \dots \varphi(\cdot - \gamma_d^i) = \sum_{i=1}^n a_i \Phi(\cdot - \gamma^i)$ for $\gamma^1, \dots, \gamma^n \in \mathbb{Z}^d$. This proves that Φ together with the linear spaces U_j generates an MRA of $L^2(\lambda^d)$. It remains to prove that the wavelets generate an orthonormal basis of $L^2(\lambda^d)$.

For an index $k \in \times_{i=1}^d \{0, 1\}$ define $a_l^{k_i}$ by $\sqrt{2}h_l$ if $k_i = 0$ and $\sqrt{2}g_l$ if $k_i = 1$ for $i = 1, \dots, d$. Furthermore, put $a_k(\gamma) := a_{\gamma_1}^{k_1} \dots a_{\gamma_d}^{k_d}$. Then, the scaling function and the wavelet generators satisfy

$$\Psi_k = \sum_{\gamma_1, \dots, \gamma_d} a_{\gamma_1}^{k_1} \dots a_{\gamma_d}^{k_d} \varphi(2 \cdot - \gamma_1) \otimes \dots \otimes \varphi(2 \cdot - \gamma_d) = \sum_{\gamma} a_k(\gamma) \Phi(M \cdot - \gamma).$$

Since φ is a scaling function, the coefficients $a_0(\gamma)$ of the scaling function Φ satisfy the relation

$$\sum_{\gamma} a_0(\gamma) = 2^{d/2} \sum_{\gamma_1, \dots, \gamma_d} h_{\gamma_1} \dots h_{\gamma_d} = 2^{d/2} \left(\sum_{\gamma_1} h_{\gamma_1} \right)^d = 2^d.$$

Furthermore, for $j, k \in \{0, 1\}^d$ and $\gamma \in \Gamma$ we have,

$$\sum_{\gamma'} a_j(\gamma') a_k(M\gamma + \gamma') = \left\{ \sum_{\gamma'_1} a_{\gamma'_1}^{j_1} a_{2\gamma_1 + \gamma'_1}^{k_1} \right\} \dots \left\{ \sum_{\gamma'_d} a_{\gamma'_d}^{j_d} a_{2\gamma_d + \gamma'_d}^{k_d} \right\} = 2^d \delta_{j,k} \delta_{\gamma,0}.$$

Indeed, we have for $s = 1, \dots, d$ and $z := \gamma_s$

$$\sum_{\gamma'_s} a_{\gamma'_s}^{j_s} a_{2\gamma_s + \gamma'_s}^{k_s} = \begin{cases} 2 \sum_l h_l g_{2z+l} & \text{if } j_s = 0 \text{ and } k_s = 1, \\ 2 \sum_l h_l h_{2z+l} & \text{if } j_s = k_s = 0, \\ 2 \sum_l g_l h_{2z+l} & \text{if } j_s = 1 \text{ and } k_s = 0, \\ 2 \sum_l g_l g_{2z+l} & \text{if } j_s = k_s = 1. \end{cases}$$

Since, the $\varphi(\cdot - z)$ form an ONB of U'_0 we have

$$\delta_{z,0} = \int_{\mathbb{R}} \varphi(x - z) \varphi(x) \, dx = \sum_{l,m} h_l h_m \delta_{2z+l,m} = \sum_l h_l h_{2z+l}.$$

In the same way,

$$\delta_{z,0} = \int_{\mathbb{R}} \psi(x - z) \psi(x) \, dx = \sum_{l,m} g_l g_m \delta_{2z+l,m} = \sum_l g_l g_{2z+l}.$$

In addition, since $U'_1 = U'_0 \otimes W'_0$ we get

$$0 = \int_{\mathbb{R}} \psi(x - z) \varphi(x) \, dx = \sum_{l,m} g_l h_m \delta_{2z+l,m} = \sum_l g_l h_{2z+l} = \sum_l g_{l-2z} h_l,$$

for all $z \in \mathbb{Z}$. Hence, the conditions of Theorem 2.2 (Theorem 1.7 in Benedetto [1993]) are fulfilled and the family of functions $\{|M|^{j/2} \Psi_k(M^j \cdot - \gamma) : \gamma \in \Gamma, k = 1, \dots, |M| - 1\}$ forms an ONB of W_j and $L^2(\lambda^d) = \oplus_{j \in \mathbb{Z}} W_j$. This finishes the proof. \square

Proof of Theorem 2.4. If $\cup_{j \in \mathbb{Z}} U_j$ is not dense in $L^p(\mu)$, there exists a $0 \neq g \in L^q(\mu)$ which fulfills $\int_{\mathbb{R}^d} f g \, d\mu = 0$ for all $f \in \overline{\cup_{j \in \mathbb{Z}} U_j}$ where q is Hölder conjugate to p . We show that the Fourier transform of g is zero which contradicts the assumption that $g \neq 0$ and hence proves that $\cup_{j \in \mathbb{Z}} U_j$ is dense. Indeed, consider the Fourier transform of this element g which we define here for reasons of simplicity as

$$\mathcal{F}g : \mathbb{R} \rightarrow \mathbb{C}, \xi \mapsto \int_{\mathbb{R}^d} g(x) e^{i\langle x, \xi \rangle} \mu(dx).$$

Since the scaling function Φ is of the form $\Phi = \otimes_{i=1}^d \varphi$ and φ is a one dimensional scaling function, we can assume that the support of Φ is contained in the cube $[0, A]^d$ for some $A \in \mathbb{N}_+$. cf. Blatter [2013]. Choose $1 > \varepsilon > 0$ arbitrary, there is a $n \in \mathbb{N}$ such that for $Q := [-An, An]^d$ we have

$$\mu(\mathbb{R}^d \setminus Q)^{1/p} < \frac{\varepsilon}{3 \cdot 2^{d-1} \max(\|g\|_{L^q(\mu)}, 1)}.$$

Fix $\xi \in \mathbb{R}^d$ arbitrary, then we get by the choice of g that

$$\begin{aligned} |\mathcal{F}g(\xi)| &\leq \left| \int_{\mathbb{R}^d} (\cos\langle x, \xi \rangle - F_1(x)) g(x) \mu(dx) \right| \\ &\quad + \left| \int_{\mathbb{R}^d} (\sin\langle x, \xi \rangle - F_2(x)) g(x) \mu(dx) \right| \end{aligned} \quad (4.7)$$

for all $F_1, F_2 \in \overline{\cup_{j \in \mathbb{Z}} U_j}$. We show that the first term in equation (4.7) is smaller than ε for suitable $F \in \overline{\cup_{j \in \mathbb{Z}} U_j}$; the second term can be treated in the same way. Therefore, we use several times the trigonometric identities $\sin = -\cos(\cdot + \frac{\pi}{2})$, as well as, $\cos(\alpha + \beta) = \cos \alpha \cos \beta - \sin \alpha \sin \beta$: we can split $\cos\langle \cdot, \xi \rangle$ in 2^{d-1} terms as $\cos\langle x, \xi \rangle = \sum_{i=1}^{2^{d-1}} b_i \cos(\xi_1 x_1 + a_{i,1}) \cdot \dots \cdot \cos(\xi_d x_d + a_{i,d})$, where the b_i are in $\{-1, 1\}$. First, we prove that each of the functions $\cos(\xi_k \cdot + a_{i,k})$ can be uniformly approximated on finite intervals. Indeed, define the kernel

$$K : \mathbb{R}^2 \rightarrow \mathbb{R}, (x, y) \mapsto \sum_{k \in \mathbb{Z}} \varphi(x - k) \varphi(y - k)$$

and for $j \in \mathbb{Z}$ the associated linear wavelet projection operator K_j as

$$K_j : L^2(\lambda) \rightarrow \overline{U_j}, \quad f \mapsto \sum_{k \in \mathbb{Z}} \langle f, 2^{j/2} \varphi(2^j \cdot - k) \rangle 2^{j/2} \varphi(2^j \cdot - k).$$

Then, K fulfills the moment condition $M(N)$ from Härdle et al. [2012] for $N = 0$: since φ is a scaling function, we have $\int_{\mathbb{R}} K(\cdot, y) dy = \sum_{k \in \mathbb{Z}} \varphi(\cdot - k) \equiv 1$. Furthermore,

$$|K(x, y)| = \left| \sum_{k \in \mathbb{Z}} \varphi(x - k) \varphi(y - k) \right| \leq (A + 1) \|\varphi\|_{\infty}^2 1_{\{|x-y| \leq A\}} =: F(x - y),$$

where we assume w.l.o.g. that $\overline{\text{supp } \varphi} \subseteq [0, A]$. Thus, F is integrable $[\lambda]$ and K satisfies the moment condition $M(0)$. Next, let $I(i, k) \supseteq [-An, An]$ be a finite interval such that $\cos(\xi_k \cdot + a_{i,k})$ is zero at the boundary of $I(i, k)$. Then by Theorem 8.1 and Remark 8.4 in Härdle et al. [2012] the uniformly continuous restriction $\cos(\xi_k \cdot + a_{i,k}) 1_{I(i,k)}$ can be approximated in $L^\infty(\lambda)$ with elements from some U_j , i.e.,

$$\|\cos(\xi_k \cdot + a_{i,k}) 1_{I(i,k)} - K_j \cos(\xi_k \cdot + a_{i,k}) 1_{I(i,k)}\|_{L^\infty(\lambda)} \rightarrow 0.$$

Thus, for $\tilde{\varepsilon} > 0$ we can choose for each factor $\cos(\xi_k \cdot + a_{i,k}) 1_{I(i,k)}$ an approximation $f_{i,k}$ in some U_j such that $\|\cos(\xi_k \cdot + a_{i,k}) 1_{I(i,k)} - f_{i,k}\|_{L^\infty(\lambda)} \leq \tilde{\varepsilon}$. This implies that for each of the $i = 1, \dots, 2^{d-1}$ products we have

$$\begin{aligned} & \|\cos(\xi_1 x_1 + a_{i,1}) 1_{I(i,1)} \cdot \dots \cdot \cos(\xi_d x_d + a_{i,d}) 1_{I(i,d)} - f_{i,1} \otimes \dots \otimes f_{i,d}\|_{L^\infty(\lambda)} \\ & \leq (1 + \tilde{\varepsilon})^d - 1 \leq d \tilde{\varepsilon} e^{d \tilde{\varepsilon}} \leq (de^d) \tilde{\varepsilon}, \end{aligned} \quad (4.8)$$

i.e., the d -dimensional approximation follows from the one dimensional approximations. Put now $F_1 := \sum_{i=1}^{2^{d-1}} b_i f_{i,1} \otimes \dots \otimes f_{i,d}$ and $\tilde{\varepsilon} := \varepsilon / (3 \cdot 2^{d-1} de^d \|g\|_{L^q(\mu)})$, then we arrive at

$$\begin{aligned} & \left| \int_{\mathbb{R}^d} (\cos \langle x, \xi \rangle - F_1(x)) g(x) \mu(dx) \right| \\ & \leq \int_Q |\cos \langle x, \xi \rangle - F_1(x)| |g(x)| \mu(dx) + \int_{\mathbb{R}^d \setminus Q} |\cos \langle x, \xi \rangle - F_1(x)| |g(x)| \mu(dx) \end{aligned} \quad (4.9)$$

We consider the terms in (4.9) separately. We can estimate the first term as follows

$$\begin{aligned} & \int_Q |\cos \langle x, \xi \rangle - F_1(x)| |g(x)| \mu(dx) \\ & \leq \sum_{i=1}^{2^{d-1}} \int_Q (de^d) \tilde{\varepsilon} |g(x)| \mu(dx) \leq 2^{d-1} de^d \|g\|_{L^q(\mu)} \tilde{\varepsilon} = \frac{\varepsilon}{3}. \end{aligned} \quad (4.10)$$

Likewise, for the second term we infer that

$$\begin{aligned} & \int_{\mathbb{R}^d \setminus B} |\cos \langle x, \xi \rangle - F_1(x)| |g(x)| \mu(dx) \\ & \leq \sum_{i=1}^{2^{d-1}} \int_{\mathbb{R}^d \setminus B} \left| \left(\prod_{k=1}^d \cos(\xi_k x_k + a_{i,k}) \right) 1_{\times_{k=1}^d I(i,k)} - \prod_{k=1}^d f_{i,k}(x_k) \right| |g(x)| \mu(dx) + \dots \\ & \quad \dots + \sum_{i=1}^{2^{d-1}} \int_{\mathbb{R}^d \setminus B} \left| \left(\prod_{k=1}^d \cos(\xi_k x_k + a_{i,k}) \right) 1_{\mathbb{R}^d \setminus \times_{k=1}^d I(i,k)} \right| |g(x)| \mu(dx) \\ & \leq 2^{d-1} de^d \tilde{\varepsilon} \|g\|_{L^q(\mu)} \mu(\mathbb{R}^d \setminus B)^{\frac{1}{p}} + 2^{d-1} \|g\|_{L^q(\mu)} \mu(\mathbb{R}^d \setminus B)^{\frac{1}{p}} \\ & = \frac{\varepsilon}{3} \cdot \frac{\varepsilon}{3 \cdot 2^{d-1} \max(\|g\|_{L^q(\mu)}, 1)} + \frac{\varepsilon}{3}. \end{aligned} \quad (4.11)$$

All in all, we have when combining equations (4.10) and (4.11) that (4.9) is less than ε as desired. \square

Proof of Theorem 2.5 and of Theorem 2.6. Throughout the proof we sometimes suppress the dependence of j from k . We prove that $\inf_{f \in \mathcal{F}_k, \|f\|_\infty \leq \beta_k} \int_{\mathbb{R}^d} |f - m|^2 d\mu_X \rightarrow 0$. Let $\varepsilon > 0$. Since $\cup_{j \in \mathbb{N}} U_j$ is dense in $L^2(\mu_X)$ there is a function f and a $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$, we have $f \in U_{j(k)}$ and $\int_{\mathbb{R}^d} |f - m|^2 d\mu_X < \varepsilon/4$. For each resolution $j(k)$ we can write

$$f = \sum_{\gamma \in K_k} a_{k,\gamma} \Psi_{j(k),\gamma} + \sum_{\gamma \notin K_k} a_{k,\gamma} \Psi_{j(k),\gamma}$$

for coefficients $a_{k,\gamma} \in \mathbb{R}$. Put $g_k := \sum_{\gamma \notin K_k} a_{k,\gamma} \Psi_{j(k),\gamma}$. The support of the g_k decreases monotonically to zero:

$$\begin{aligned} \{g_k \neq 0\} & \subseteq \{x \in \mathbb{R}^d : M^j x - \gamma \in [0, L]^d, \|\gamma\|_\infty > w_k\} \\ & \subseteq \{x \in \mathbb{R}^d : \|M^j x\|_\infty \geq \|\gamma\|_\infty - L, \|\gamma\|_\infty > w_k\} \\ & \subseteq \{x \in \mathbb{R}^d : \|M^j x\|_2 \geq w_k - L\} \\ & \subseteq \{x \in \mathbb{R}^d : \|S^{-1}\|_2 (\lambda_{\max})^j \|S\|_2 \|x\|_2 \geq w_k - L\} \downarrow \emptyset \quad (k \rightarrow \infty), \end{aligned}$$

by the assumption that $(\lambda_{\max})^{j(k)}/w_k \rightarrow 0$ as $k \rightarrow \infty$. Furthermore, there is a $k_1 \in \mathbb{N}$ such that for all $k \geq k_1$ we have $\int_{\mathbb{R}^d} f^2 1\{\mathbb{R}^d \setminus [-k_1, k_1]^d\} d\mu_X < \varepsilon/4$. Hence, there is a $k_2 \in \mathbb{N}$ such that both

$$[-k_1, k_1]^d \subseteq \cup_{\gamma \in K_k} \text{supp } \Psi_{j(k), \gamma} \text{ and } \|f 1\{[-k_1, k_1]^d\}\|_\infty \leq \beta_k$$

for all $k \geq k_2$. In particular, $f 1\{[-k_1, k_1]^d\}$ is eligible in that it is in $T_{\beta_k} \mathcal{F}_k$ and $\int_{\mathbb{R}^d} |m - f 1\{[-k_1, k_1]^d\}|^2 d\mu_X < \varepsilon$ as desired. For the second part, we merely need to perform the same computations as in the proof of Theorem 1.7. It remains to compute $\kappa_k(\varepsilon, \beta_k) := \log H_{T_{\beta_k} \mathcal{F}_k}(\varepsilon/(128\beta_k))$. We use the bound given in Proposition A.1

$$H_{T_{\beta_k} \mathcal{F}_k} \leq 3 \exp \left\{ 2((2w_k + 1)^d + 1) \log(768e\beta_k^2/\varepsilon) \right\}, \text{ i.e., } \kappa_k(\varepsilon, \beta_k) \in O(w_k^d \log(\beta_k))$$

for $\varepsilon > 0$ fix. The estimator is weakly consistent if

$$w_k^d \beta_k^2 \log \beta_k \prod_{i=1}^N \log n_i(k) / \left(\prod_{i=1}^N n_i(k) \right)^{\rho - N/(N+1)} \rightarrow 0 \text{ as } k \rightarrow \infty.$$

Furthermore, again with Theorem 1.7 and for the case of a full lattice, if additionally

$$\beta_k^2 (\log k)^{1+\delta} \prod_{i=1}^N \log n_i(k) / \left(\prod_{i=1}^N n_i(k) \right)^{1/(N+1)} \rightarrow 0 \text{ as } k \rightarrow \infty$$

for some $\delta > 0$, the estimator is strongly consistent. The statement which concerns the rate of convergence follows immediately from Theorem 1.9. \square

APPENDIX A. EXPONENTIAL INEQUALITIES FOR DEPENDENT SUMS

In this section, we give a short review on important concepts which we shall use throughout this article. We start with a proposition on the covering number. Denote by $\mathcal{G}^+ := \{(z, t) \in \mathbb{R}^d \times \mathbb{R} : t \leq g(z) : g \in \mathcal{G}\}$ the class of all subgraphs of the class \mathcal{G} . Condition 1.6 is satisfied if the Vapnik-Chervonenkis dimension of \mathcal{G}^+ is at least two, i.e., $\mathcal{V}_{\mathcal{G}^+} \geq 2$ and if ε sufficiently small:

Proposition A.1 (Bound on the covering number, Györfi et al. [2002] Theorem 9.4, Haussler [1992]). *Let $[a, b] \subset \mathbb{R}$ be a finite interval. Let \mathcal{G} be a class of uniformly bounded real valued functions $g : \mathbb{R}^d \mapsto [a, b]$ such that $\mathcal{V}_{\mathcal{G}^+} \geq 2$. Let $0 < \varepsilon < (b - a)/4$. Then for any probability measure ν on $\mathcal{B}(\mathbb{R}^d)$*

$$N(\varepsilon, \mathcal{G}, \|\cdot\|_{L^p(\nu)}) \leq 3 \left(\frac{2e(b-a)^p}{\varepsilon^p} \log \frac{3e(b-a)^p}{\varepsilon^p} \right)^{\mathcal{V}_{\mathcal{G}^+}}.$$

In particular, in the case that \mathcal{G} is an r -dimensional linear space, we have $\mathcal{V}_{\mathcal{G}^+} \leq r + 1$.

Davydov's inequality relates the covariance of two random variables to the α -mixing coefficient:

Proposition A.2 (Davydov's inequality). *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $\mathcal{G}, \mathcal{H} \subseteq \mathcal{A}$ be sub- σ -algebras. Denote by $\alpha := \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \mathcal{G}, B \in \mathcal{H}\}$ the α -mixing coefficient of \mathcal{G} and \mathcal{H} . Let $p, q, r \geq 1$ be Hölder conjugate, i.e., $p^{-1} + q^{-1} + r^{-1} = 1$. Let ξ (resp. η) be in $L^p(\mathbb{P})$ and \mathcal{G} -measurable (resp. in $L^q(\mathbb{P})$ and \mathcal{H} -measurable). Then $|\text{Cov}(\xi, \eta)| \leq 10 \alpha^{1/r} \|\xi\|_{L^p(\mathbb{P})} \|\eta\|_{L^q(\mathbb{P})}$.*

The aim of this section is to derive upper bounds on the probability of events of the type

$$\left\{ \sup_{g \in \mathcal{G}} \left| \frac{1}{|I_n|} \sum_{s \in I_n} g(Z(s)) - \mathbb{E}[g(Z(e_N))] \right| > \varepsilon \right\}, \quad (\text{A.1})$$

for a given class of functions \mathcal{G} , a random field $\{Z(s) : s \in \mathbb{Z}^N\}$ and subsets $I_n \subseteq \mathbb{Z}^N$. Since equation (A.1) is a priori not an event, we shall assume throughout the paper that the classes \mathcal{G} are sufficiently regular and that (A.1) is \mathcal{A} -measurable.

The next theorem is crucial for the analysis in Sections 1 and 2; we give a modified version of the N -dimensional Bernstein inequality from Valenzuela-Domínguez and Franke [2005] which holds even for non-stationary random fields of the type $\{Z(s) : s \in I\}$ under some weaker regularity conditions.

Theorem A.3 (Bernstein inequality for strong spatial mixing, Valenzuela-Domínguez and Franke [2005]). *Let $Z := \{Z(s) : s \in I\}$ be a real-valued random field defined on a subset of the N -dimensional lattice \mathbb{Z}^N . Let Z be strong mixing with mixing coefficients $\{\alpha_k : k \in \mathbb{N}_+\}$ such that each $Z(s)$ is bounded by a uniform*

constant B and has expectation zero and the variance of $Z(s)$ is uniformly bounded by σ^2 . Furthermore, put $\bar{\alpha}_k := \sum_{u=1}^k u^N \alpha_u$. Then for all $\varepsilon > 0$ and $\beta > 0$ such that $0 < 2^{N+1} B \tilde{P} e \beta < 1$

$$\mathbb{P} \left(\left| \sum_{s \in I_n} Z(s) \right| > \varepsilon \right) \leq 2 \exp \left\{ D_1 \sqrt{e} 2^N \frac{\tilde{n}}{\tilde{P}} \alpha_q^{\tilde{P}/[\tilde{n}(2^N+1)]} \right\} \cdot \exp \left\{ -\beta \varepsilon + 2^{3N} \beta^2 e (\sigma^2 + 4D_2 B^2 \bar{\alpha}_P) \tilde{n} \right\}, \quad (\text{A.2})$$

where $D_1, D_2 > 0$ are constants depending on the dimension N and $P(n), Q(n)$ are arbitrary non-decreasing sequences in \mathbb{N}_+^N satisfying for each $1 \leq i \leq N$

$$\begin{aligned} 1 &\leq Q_i(n_i) \leq P_i(n_i) < Q_i(n_i) + P_i(n_i) < n_i \text{ and} \\ \tilde{n} &:= n_1 \cdot \dots \cdot n_N, \quad \tilde{P} := P_1(n_1) \cdot \dots \cdot P_N(n_N) \\ q &:= \min \{Q_1(n_1), \dots, Q_N(n_N)\}, \quad P := \max \{P_1(n_1), \dots, P_N(n_N)\}. \end{aligned}$$

To conclude this section, we state useful technical results based on Theorem A.3.

Proposition A.4. *Let the real valued random field Z satisfy Condition 1.4. The $Z(s)$ have expectation zero and are bounded by B . There are constants $A_1, A_2 \in \mathbb{R}_+$ which depend on the lattice dimension N and on the bound of the mixing coefficients which is determined by the numbers c_0 and c_1 but not on $n \in \mathbb{N}_+^N$ and not on B such that for all $n \in \mathbb{N}_+^N$ with $\min_{1 \leq i \leq N} n_i \geq e^2$ and $\varepsilon > 0$*

$$\mathbb{P} \left(\left| \sum_{s \in I_n} Z(s) \right| > \varepsilon \right) \leq A_1 \exp \left(-A_2 \varepsilon B^{-1} \left(\prod_{i=1}^N n_i \right)^{-N/(N+1)} \left(\prod_{i=1}^N \log n_i \right)^{-1} \right).$$

Proof of Proposition A.4. We make the definitions: $P_i(n_i) := Q_i(n_i) := \lfloor n_i^{N/(N+1)} \log n_i \rfloor$ for $i = 1, \dots, N$. Furthermore, we denote the smallest coordinate of $n \in \mathbb{N}^N$ by $n^* := \min_{1 \leq i \leq N} n_i$. We consider the first factor of the RHS of (A.2) and show that under the stated conditions we have

$$\sup \left\{ \exp \left(D_1 \sqrt{e} 2^N \frac{\tilde{n}}{\tilde{P}} \alpha_q^{\tilde{P}/[\tilde{n}(2^N+1)]} \right) : n \in \mathbb{Z}^N, n^* \geq e^2 \right\} < \infty. \quad (\text{A.3})$$

By assumption the mixing coefficient satisfies $\alpha(q) \leq c_0 \exp(-c_1 q)$, for two constants $c_0, c_1 \in \mathbb{R}_{\geq 0}$ and $q = \min_{1 \leq i \leq N} Q_i$. Therefore it suffices to show that

$$\log(\tilde{n}/\tilde{P}) - c_1/(2^N + 1) q \tilde{P}/\tilde{n} \rightarrow -\infty \text{ as } n^* \rightarrow \infty. \quad (\text{A.4})$$

Note that for $a, b \geq 2$, we have $ab \geq a + b$. We make the definition $\eta := N/(N+1)$. Let $n^* \geq e^2$, then for any constant $C \in \mathbb{R}_+$

$$\begin{aligned} &\log \left(\left(\prod_{i=1}^N n_i \right)^{1-\eta} \left(\prod_{i=1}^N \log n_i \right)^{-1} \right) - C(n^*)^\eta \log n^* \left(\prod_{i=1}^N n_i \right)^{\eta-1} \left(\prod_{i=1}^N \log n_i \right) \\ &\leq (N+1)^{-1} \sum_{i=1}^N \log n_i - C(n^*)^{\eta+N(\eta-1)} \left(\log n^* \prod_{i=1}^N \log n_i \right) \\ &\leq (N+1)^{-1} \prod_{i=1}^N \log n_i - C \left(\log n^* \prod_{i=1}^N \log n_i \right) \\ &= \left((N+1)^{-1} - C \log n^* \right) \prod_{i=1}^N \log n_i \rightarrow -\infty \text{ as } n^* \rightarrow \infty. \end{aligned}$$

This proves (A.4) and consequently, that (A.3) is finite. We come to the second term inside the second factor of (A.2). Define $\beta := (2^{N+2} e B \tilde{P})^{-1}$ which fulfills the requirements of Theorem A.3. Then,

$$\begin{aligned} &\sup \left\{ 2^{3N} \beta^2 e (\sigma^2 + 4D_2 B^2 \bar{\alpha}_P) \tilde{n} : n \in \mathbb{N}^N, n^* \geq e^2 \right\} \\ &\leq \sup \left\{ 2^{3N} (2^{N+2} \tilde{P})^{-2} (1 + 4D_2 \bar{\alpha}_P) \tilde{n} : n \in \mathbb{N}^N, n^* \geq e^2 \right\} < \infty. \end{aligned}$$

This proves that $\mathbb{P} \left(\left| \sum_{s \in I_n} Z(s) \right| > \varepsilon \right) \leq A \exp \left(-\varepsilon / (2^{N+2} e B \tilde{P}) \right)$ for a constant $A \in \mathbb{R}_+$ which only depends on the lattice dimension N and on the bound of the mixing coefficients determined by the numbers c_0 and c_1 . \square

We can prove with the previous proposition an important statement

Theorem A.5 (Large deviations for strong spatial mixing data). *Let the random field Z satisfy Condition 1.4 and have equal marginal distributions. Let \mathcal{G} be a set of measurable functions $g : \mathbb{R}^d \rightarrow [0, B]$ for $B \in [1, \infty)$ which satisfies Condition 1.6. Then given that (A.1) is \mathcal{A} -measurable, for any $\varepsilon > 0$ and $n \in \mathbb{N}_+^N$ such that $\min_{1 \leq i \leq N} n_i \geq e^2$*

$$\begin{aligned} & \mathbb{P} \left(\sup_{g \in \mathcal{G}} \left| \frac{1}{|I_n|} \sum_{s \in I_n} g(Z(s)) - \mathbb{E}[g(Z(e_N))] \right| > \varepsilon \right) \\ & \leq A_1 H_{\mathcal{G}} \left(\frac{\varepsilon}{32} \right) \left\{ \exp \left(-\frac{A_2 \varepsilon^2 |I_n|}{B^2} \right) + \exp \left(-\frac{A_3 \varepsilon |I_n|}{B \left(\prod_{i=1}^N n_i \right)^{N/(N+1)} \prod_{i=1}^N \log n_i} \right) \right\} \end{aligned}$$

where A_1, A_2 and A_3 only depend on $N \in \mathbb{N}_+$ and on the bound of the mixing coefficients given by $c_0, c_1 \in \mathbb{R}_+$.

In practice, we use the bound given in Theorem A.5 on an increasing sequence $(n(k) : k \in \mathbb{N}) \subseteq \mathbb{Z}^N$ and on increasing function classes \mathcal{G}_k whose essential bounds B_k increase with the size of the index sets $I_{n(k)}$. Hence, it is possible to omit the first $|I_n|$ -dependent term in the above theorem under a certain condition: let a sequence of function classes \mathcal{G}_k with bounds B_k and a sequence $(\varepsilon_k : k \in \mathbb{N}_+) \subseteq \mathbb{R}_+$ be given such that

$$\lim_{k \rightarrow \infty} \varepsilon_k |I_{n(k)}| \left\{ B_k \left(\prod_{i=1}^N n_i(k) \right)^{N/(N+1)} \prod_{i=1}^N \log n_i(k) \right\} = \infty,$$

then the above equation reduces to

$$\begin{aligned} & \mathbb{P} \left(\sup_{g \in \mathcal{G}_k} \left| \frac{1}{|I_{n(k)}|} \sum_{s \in I_{n(k)}} g(Z(s)) - \mathbb{E}[g(Z(e_N))] \right| > \varepsilon_k \right) \\ & \leq A_1 H_{\mathcal{G}_k} \left(\frac{\varepsilon_k}{32} \right) \exp \left(-\frac{A_2 \varepsilon_k |I_{n(k)}|}{B_k \left(\prod_{i=1}^N n_i(k) \right)^{N/(N+1)} \prod_{i=1}^N \log n_i(k)} \right) \end{aligned}$$

with new constants $A_1, A_2 \in \mathbb{R}_+$.

Proof of Theorem A.5. We assume that the probability space is additionally endowed with the i.i.d. random variables $Z'(s)$ for $s \in I_n$ which have the same marginal laws as the $Z(s)$. We define

$$S_n(g) := \frac{1}{|I_n|} \sum_{s \in I_n} g(Z(s)) \text{ and } S'_n(g) := \frac{1}{|I_n|} \sum_{s \in I_n} g(Z'(s)).$$

Thus, we can decompose

$$\begin{aligned} & \mathbb{P} \left(\sup_{g \in \mathcal{G}} |S_n(g) - \mathbb{E}[g(Z(e_N))]| > \varepsilon \right) \\ & \leq \mathbb{P} \left(\sup_{g \in \mathcal{G}} |S_n(g) - S'_n(g)| > \frac{\varepsilon}{2} \right) + \mathbb{P} \left(\sup_{g \in \mathcal{G}} |S'_n(g) - \mathbb{E}[g(Z'(e_N))]| > \frac{\varepsilon}{2} \right) \end{aligned} \quad (\text{A.5})$$

and apply Theorem 9.1 from Györfi et al. [2002] to second term on the right-hand side of (A.5) which is bounded by

$$\mathbb{P} \left(\sup_{g \in \mathcal{G}} |S'_n(g) - \mathbb{E}[g(Z'(e_N))]| > \frac{\varepsilon}{2} \right) \leq 8 H_{\mathcal{G}} \left(\frac{\varepsilon}{16} \right) \exp \left(-\frac{|I_n| \varepsilon^2}{512 B^2} \right). \quad (\text{A.6})$$

To get a bound on the first term of the right-hand side of (A.5), we apply for fix $\omega \in \Omega$ the Condition 1.6 to the set $\{Z(s, \omega), Z'(s, \omega) : s \in I_n\}$. Let $g_k^*(\omega)$ for $k = 1, \dots, H^* := H_{\mathcal{G}} \left(\frac{\varepsilon}{32} \right)$ be chosen as in Condition 1.6, possibly with some redundant $g_k^*(\omega)$ for $\tilde{H}(\omega) < k \leq H^*$ where $\tilde{H}(\omega)$ is the number of non-redundant functions. Note that H^* is deterministic. Define the random sets for $k = 1, \dots, H^*$ by

$$\begin{aligned} U_k(\omega) := \left\{ g \in \mathcal{G} : \frac{1}{2|I_n|} \sum_{s \in I_n} |g(Z(s, \omega)) - g_k^*(Z(s, \omega))| \right. \\ \left. + |g(Z'(s, \omega)) - g_k^*(Z'(s, \omega))| < \frac{\varepsilon}{32} \right\}, \end{aligned}$$

note that some $U_k(\omega)$ might be redundant for $\tilde{H}(\omega) < k \leq H^*$. This implies that for each $\omega \in \Omega$ we can write $\mathcal{G} = U_1(\omega) \cup \dots \cup U_k(\omega)$, consequently,

$$\begin{aligned} \mathbb{P} \left(\sup_{g \in \mathcal{G}} |S_n(g) - S'_n(g)| > \frac{\varepsilon}{2} \right) &= \mathbb{P} \left(\max_{1 \leq k \leq H^*} \sup_{g \in U_k} |S_n(g) - S'_n(g)| > \frac{\varepsilon}{2} \right) \\ &\leq \mathbb{E} \left[\sum_{k=1}^{\tilde{H}} 1_{\left\{ \sup_{g \in U_k} |S_n(g) - S'_n(g)| > \frac{\varepsilon}{2} \right\}} \right] \leq \sum_{k=1}^{H^*} \mathbb{P} \left(\sup_{g \in U_k} |S_n(g) - S'_n(g)| > \frac{\varepsilon}{2} \right). \end{aligned} \quad (\text{A.7})$$

In the following, we suppress the ω -wise notation; let now $g \in U_k$ be arbitrary but fixed, then

$$|S_n(g) - S'_n(g)| \leq 2 \frac{\varepsilon}{32} + |S_n(g_k^*) - S'_n(g_k^*)|. \quad (\text{A.8})$$

Thus, using equation (A.8), we get for each summand in (A.7)

$$\begin{aligned} \mathbb{P} \left(\sup_{g \in U_k} |S_n(g) - S'_n(g)| > \frac{\varepsilon}{2} \right) &\leq \mathbb{P} \left(|S_n(g_k^*) - S'_n(g_k^*)| > \frac{7\varepsilon}{16} \right) \\ &\leq \mathbb{P} \left(\left| S_n(g_k^*) - \mathbb{E} [g_k^*(Z(e_N))] \right| > \frac{7\varepsilon}{32} \right) + \mathbb{P} \left(\left| S'_n(g_k^*) - \mathbb{E} [g_k^*(Z'(e_N))] \right| > \frac{7\varepsilon}{32} \right). \end{aligned} \quad (\text{A.9})$$

The second term on the right-hand side of (A.9) can be estimated using Hoeffding's inequality, we have

$$\mathbb{P} \left(\left| S'_n(g_k^*) - \mathbb{E} [g_k^*(Z'(e_N))] \right| > \frac{7\varepsilon}{32} \right) \leq 2 \exp \left\{ -\frac{98 |I_n| \varepsilon^2}{32^2 B^2} \right\}. \quad (\text{A.10})$$

We apply the Bernstein inequality for strong spatial mixing data from Theorem A.3 to the first term of equation (A.9). We obtain for the first term on the right-hand side of (A.9) with Proposition A.4

$$\mathbb{P} \left(\left| S_n(g_k^*) - \mathbb{E} [g_k^*(Z(e_N))] \right| > \frac{7\varepsilon}{32} \right) \leq 2A_1 \exp \left\{ -\frac{A_2 \varepsilon |I_n|}{B \left(\prod_{i=1}^N n_i \right)^{N/(N+1)} \prod_{i=1}^N \log n_i} \right\}. \quad (\text{A.11})$$

And all in all, using that $H_{\mathcal{G}} \left(\frac{\varepsilon}{16} \right) \leq H_{\mathcal{G}} \left(\frac{\varepsilon}{32} \right)$ and with the help of equation (A.6), and equations (A.10) and (A.11) plugged in (A.9) and that again in (A.7) we get the result - using the notation $\tilde{n} = \prod_{i=1}^N n_i$

$$\begin{aligned} &\mathbb{P} \left(\sup_{g \in \mathcal{G}} \left| \frac{1}{|I_n|} \sum_{s \in I_n} g(Z(s)) - \mathbb{E} [g(Z(e_N))] \right| > \varepsilon \right) \\ &\leq 8H_{\mathcal{G}} \left(\frac{\varepsilon}{16} \right) \exp \left(-\frac{\varepsilon^2 |I_n|}{512 B^2} \right) \\ &\quad + 2H_{\mathcal{G}} \left(\frac{\varepsilon}{32} \right) \left\{ \exp \left(-\frac{98 \varepsilon^2 |I_n|}{32^2 B^2} \right) + A_1 \exp \left(-\frac{A_2 \varepsilon |I_n|}{B \tilde{n}^{N/(N+1)} \prod_{i=1}^N \log n_i} \right) \right\} \\ &\leq (10 + 2A_1) H_{\mathcal{G}} \left(\frac{\varepsilon}{32} \right) \left\{ \exp \left(-\frac{\varepsilon^2 |I_n|}{512 B^2} \right) + \exp \left(-\frac{A_2 \varepsilon |I_n|}{B \tilde{n}^{N/(N+1)} \prod_{i=1}^N \log n_i} \right) \right\}. \end{aligned}$$

This finishes the proof. \square

APPENDIX B. ERGODIC THEORY FOR SPATIAL PROCESSES

In this section, we give a review on important concepts of ergodicity when dealing with random fields on subgroups of the discrete group \mathbb{Z}^N . For further reading consult Tempelman [2010].

Definition B.1 (Dynamical systems and ergodicity). Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(G, +)$ a locally compact, abelian Hausdorff group which fulfills the second axiom of countability. We write for $x, y \in G$ arbitrary $x - y$ for $x + (-y)$ and $-y$ is the $+$ -inverse of y . Furthermore, let ν be a Haar measure on $\mathcal{B}(G)$, i.e., for all $x \in G$ and for all Borel sets $B \in \mathcal{B}(G)$ we have $\nu(B) = \nu(x + B)$.

A family of bijective mappings $\{T_x : \Omega \rightarrow \Omega, x \in G\}$ is called a flow if it fulfills the following three conditions

- (1) T_x is measure-preserving, i.e., $\mathbb{P}(A) = \mathbb{P}(T_x A)$ for all $A \in \mathcal{A}$ and for all $x \in G$,
- (2) $T_{x+x'} = T_x \circ T_{x'}$ and $T_x \circ T_{-x} = Id_{\Omega}$ for all $x, x' \in G$,
- (3) the map $G \times \Omega \ni (x, \omega) \mapsto T_x \omega$ is $\mathcal{B}(G) \otimes \mathcal{A} - \mathcal{A}$ -measurable.

Let $T = \{T_x : x \in G\}$ be a flow in $(\Omega, \mathcal{A}, \mathbb{P})$, then the quadruple $(\Omega, \mathcal{A}, \mathbb{P}, T)$ is called a *dynamical system*. The dynamical system is called ergodic if the invariant σ -field $\mathcal{I} := \{A \in \mathcal{A} : A = T_x A \forall x \in G\}$ is \mathbb{P} -trivial, i.e., if for all $A \in \mathcal{I}$ we have $\mathbb{P}(A) \in \{0, 1\}$.

Let now $\Gamma \leq \mathbb{Z}^N$ be a subgroup and $Z = \{Z(s) : s \in \Gamma\}$ be a stationary random field on $(\Omega, \mathcal{A}, \mathbb{P})$ where each $Z(s)$ takes values in the measure space (S, \mathfrak{S}) . Let ν be the counting measure on $\mathcal{B}(\Gamma)$. Put $\mathbb{P}_Z := \mathbb{P}_{\{Z(s) : s \in \Gamma\}}$ for the probability measure on $\otimes_{s \in \Gamma} \mathfrak{S}$ induced by the finite dimensional distributions of Z and define on the path space $(\times_{s \in \Gamma} S, \otimes_{s \in \Gamma} \mathfrak{S}, \mathbb{P}_Z)$ the family of translations

$$T_t : \times_{s \in \Gamma} S \rightarrow \times_{s \in \Gamma} S, (z(s) : s \in \Gamma) \mapsto (z(s+t) : s \in \Gamma) \quad \text{for } t \in \Gamma,$$

which is a flow because Z is stationary. Then Z is called ergodic if and only if the quadruple $(\times_{s \in \Gamma} S, \otimes_{s \in \Gamma} \mathfrak{S}, \mathbb{P}_Z, T)$ is ergodic.

The next result is an extension of Birkhoff's celebrated ergodic theorem it can be found in Tempelman [2010]

Theorem B.2 (Ergodic theorem, Tempelman [2010]). *Let $(\Omega, \mathcal{A}, \mathbb{P}, T)$ be a dynamical system. Furthermore, let $\{W_n : n \in \mathbb{N}\} \subseteq G$ be an increasing sequence of Borel sets of G such that $0 < \nu(W_n) < \infty$ for all $n \in \mathbb{N}$ which fulfills both*

$$\lim_{n \rightarrow \infty} \frac{\nu(W_n \cap (W_n - x))}{\nu(W_n)} = 1 \text{ for all } x \in G \text{ and } \sup_{n \geq 0} \frac{\nu(W_n - W_n)}{\nu(W_n)} < \infty,$$

where $W_n - W_n := \{x - y : x, y \in W_n\}$. Then, for an integrable random variable $X \in L^1(\mathbb{P})$

$$\lim_{n \rightarrow \infty} \frac{1}{\nu(W_n)} \int_{W_n} X(T_x \omega) \nu(dx) = \mathbb{E}[X | \mathcal{I}](\omega) \quad \text{for } \mathbb{P}\text{-almost every } \omega \in \Omega.$$

Proof. Compare Tempelman [2010] Chapter 6, in particular Proposition 1.3 and Corollary 3.2. \square

We are now prepared to state a well-known and useful result, cf. Hannan [2009] Theorem IV.2 and the discussion thereafter for a treatment of one-dimensional stochastic processes.

Proposition B.3 (Stationarity and mixing imply ergodicity). *Let $0 \neq \Gamma \leq \mathbb{Z}^N$ be a subgroup and let the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ be endowed with the stationary process $Z = \{Z(s) : s \in \Gamma\}$ for which each $Z(s)$ takes values in (S, \mathfrak{S}) and which fulfills the strong mixing condition from Definition 1.3. Then Z is ergodic.*

Proof. Let $A \in \mathcal{I}$ be an T -invariant set of paths of Z , it suffices to show that $\mathbb{P}(A) \in \{0, 1\}$, i.e.,

$$\mathbb{P}_Z(A) = \mathbb{P}_Z(A \cap T_x A) \rightarrow \mathbb{P}_Z(A) \mathbb{P}_Z(T_x A) = \mathbb{P}_Z(A)^2 \text{ as } x \rightarrow \infty.$$

Let $\varepsilon > 0$ be given and let $A, B \in \otimes_{k \in \Gamma} \mathfrak{S}$ be two sets of paths of Z . Then by Carathéodory's extension theorem there are $m, n \in \mathbb{Z}$ such that there are $A^m \in \otimes_{k \leq m \cdot e_N} \mathfrak{S}$ and $B^n \in \otimes_{k \geq n \cdot e_N} \mathfrak{S}$ with the property that both

$$\mathbb{P}_Z(A \Delta A^m) < \frac{\varepsilon}{5} \text{ and } \mathbb{P}_Z(B \Delta B^n) < \frac{\varepsilon}{5}.$$

Furthermore, by the strong mixing property from Definition 1.3 there is an $x^* = r \cdot e_N \in \mathbb{Z}^N$ such that for $x \geq x^*$, $x \in \Gamma$ we have

$$|\mathbb{P}_Z(A^m \cap T_x B^n) - \mathbb{P}_Z(A^m) \mathbb{P}_Z(T_x B^n)| < \frac{\varepsilon}{5}.$$

Consequently, we have for all $x \geq x^*$

$$\begin{aligned} & \left| \mathbb{P}(Z \in A, Z \in T_x B) - \mathbb{P}(Z \in A) \mathbb{P}(Z \in T_x B) \right| \\ & \leq \mathbb{P}(Z \in A \setminus A^m, Z \in T_x B) + \mathbb{P}(Z \in A^m, Z \in T_x B \setminus B^n) \\ & \quad + \left| \mathbb{P}(Z \in A^m, Z \in T_x B^n) - \mathbb{P}(Z \in A^m) \mathbb{P}(Z \in T_x B^n) \right| \\ & \quad + \mathbb{P}(Z \in A^m) \mathbb{P}(Z \in T_x B \setminus B^n) + \mathbb{P}(Z \in A \setminus A^m) \mathbb{P}(Z \in T_x B) < \varepsilon. \end{aligned}$$

\square

Next, we state a strong law of large numbers for homogeneous strong mixing random fields which we use later. We denote by $e_N := (1, \dots, 1)^T$ the N -dimensional vector whose entries are equal to 1. For an N -dimensional cube in \mathbb{Z}^N that is spanned by two points $a, b \in \mathbb{Z}^N$, we write $[a..b]$.

Theorem B.4 (Ergodicity on a lattice). *Let $0 \neq \Gamma \leq \mathbb{Z}^N$ be a nontrivial subgroup and $\{Z(s) : s \in \Gamma\}$ be a homogeneous strong mixing random field on $(\Omega, \mathcal{A}, \mathbb{P})$ for some dimension $N \in \mathbb{N}_+$. Let $(n(k) : k \in \mathbb{N}) \subseteq \mathbb{N}^N$ be an increasing sequence such that $e_N \leq n(k) \leq n(k+1)$ for which at least one coordinate converges to infinity. Then the sequence of index sets $I_{n(k)} := \{z \in \Gamma : e_N \leq z \leq n(k)\}$ is admissible in the sense of Theorem B.2. In particular, we have*

$$\frac{1}{|I_{n(k)}|} \sum_{s \in I_{n(k)}} Z(s) \rightarrow \mathbb{E}[Z(e_N)] \quad \text{a.s. as } k \rightarrow \infty.$$

Proof. Since any subgroup of \mathbb{Z}^N is isomorphic to \mathbb{Z}^u for $0 \leq u \leq N$, $u \in \mathbb{N}$, it suffices to consider the case $\Gamma = \mathbb{Z}^N$, $N \in \mathbb{N}_+$. In this case one computes easily that the regularity conditions of Theorem B.2 are satisfied. The conclusion follows then from this theorem in combination with Proposition B.3. \square

REFERENCES

- J.J. Benedetto. *Wavelets: mathematics and applications*. Studies in Advanced Mathematics. Taylor & Francis, 1993.
- C. Blatter. *Wavelets: eine Einführung*. Advanced Lectures in Mathematics. Vieweg+Teubner Verlag, 2013.
- Pierre Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Texts in Applied Mathematics. Springer, 1999.
- N.A.C. Cressie. *Statistics for spatial data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley, 1993.
- László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Berlin, New York, Heidelberg, 2002.
- E.J. Hannan. *Multiple time series*. Wiley Series in Probability and Statistics. Wiley, 2009.
- W. Härdle, G. Kerkycharian, D. Picard, and A. Tsybakov. *Wavelets, approximation, and statistical applications*. Lecture Notes in Statistics. Springer New York, 2012.
- David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.
- Onésimo Hernández-Lerma and Jean B. Lasserre. Further criteria for positive Harris recurrence of Markov chains. *Proceedings of the American Mathematical Society*, 129(5):pp. 1521–1524, 2001.
- Mark S. Kaiser, Soumendra N. Lahiri, and Daniel J. Nordman. Goodness of fit tests for a class of Markov random field models. *Ann. Statist.*, 40(1):104–130, 02 2012. doi: 10.1214/11-AOS948.
- Sean P. Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Cambridge university press, 2009.
- A.A. Tempelman. *Ergodic theorems for group actions: informational and thermodynamical Aspects*. Mathematics and Its Applications. Springer Netherlands, 2010.
- Eduardo Valenzuela-Domínguez and Jürgen Franke. A Bernstein inequality for strongly mixing spatial random processes. Technical report, Preprint series of the DFG priority program 1114 “Mathematical methods for time series analysis and digital image processing”, January 2005.
- E-mail address: krebs@mathematik.uni-kl.de